UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE RELAÇÕES INTERNACIONAIS

KING'S COLLEGE LONDON

MATHEUS SOLDI HARDT

Who, What and When: How Media and Politicians Shape the Brazilian Debate on Foreign Affairs

São Paulo
2019

MATHEUS SOLDI HARDT

Who, What and When: How Media and Politicians Shape the Brazilian Debate on Foreign Affairs

Tese apresentada ao Programa de Pós-Graduação em Relações Internacionais do Instituto de Relações Internacionais da Universidade de São Paulo, para a obtenção do duplo diploma de Doutor em Ciências junto com o King's College de Londres.

Orientador(a): Profa. Dra. Janina Onuki (USP)

Co-orientador(a): Prof. Dr. Anthony Pereira (KCL)

São Paulo

2019

Catalogação na publicação

Serviço de Biblioteca e Documentação

Instituto de Relações Internacionais da Universidade de São Paulo

Hardt, Matheus Soldi

Who, what and when: how media and politicians shape the Brazilian debate on foreign affairs / Matheus Soldi Hardt ; orientado por Janina Onuki e Anthony Pereira. – São Paulo, 2019.

183 p.

Tese (Doutorado) – Instituto de Relações Internacionais. Universidade de São Paulo, São Paulo, 2019.

1. Discursos Políticos 2. Jornais 3. Análise de Tópicos 4. LDA 5. Brasil I. Onuki, Janina, orient. II. Pereira, Anthony, orient. III. Título.

CDD – 320.014

Responsável: Giseli Adornato de Aguiar - CRB-8/6813

KING'S College LONDON

Universidade de São Paulo
Instituto de Relações Internacionais

CERTIFICATE OF DEFENSE APPROVAL FOR DOCTORAL THESIS

Certificate of public defense approval for the Doctoral Thesis of **Mr. Matheus Soldi Hardt** in the International Relations Postgraduate Program of the Institute of International Relations of the University of São Paulo (IRI-USP).

As part of the requirements to obtain a PhD degree, on the 10th of July 2019 Mr. Matheus Soldi Hardt defended his doctoral thesis entitled:

"Who, what and when: How media and politicians shape the Brazilian debate on foreign affairs"

After the public defense was opened, the president gave the floor to the candidate for his oral presentation. Following the student's presentation, the examiners questioned the candidate in accordance with defense procedures. The committee members then signed this form indicating the result:

Examiner's Name	Position	Institution	Result
Janina Onuki	President	IRI - USP	non-voting
Cristiane de Andrade Lucena Carneiro	Examiner 1	IRI - USP	aprovado
Flavio Leão Pinheiro	Examiner 2	UFABC	aprovado
Vinicius Mariano de Carvalho	Examiner 3	KCL	aprovado

Final Result: aprovado

Committee's Evaluation

Note: if the candidate is disapproved by any of the members, the completion of the committee's evaluation is mandatory

This certificate was drawn up by Giselle de Castro fcastro, Head of the Postgraduate Office, on the 19th of June 2019.

pl Janina Onuki
Cristiane de Andrade Lucena Carneiro (Examiner 1)

Flavio Leão Pinheiro (Examiner 2)

pl Janina Onuki
Vinicius Mariano de Carvalho (Examiner 3)

Janina Onuki
Janina Onuki
President of the Committee

The defense was approved by the Postgraduate Studies Committee on 08 August 2019 and, therefore, the student has received the title of PhD in Sciences, awarded by the International Relations Program.

Adriana Schor
President of the Postgraduate Studies Committee

Adriana Schor
Presidente
Comissão de Pós-Graduação
IRI-USP

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Abstract

What do politicians talk about when discussing foreign affairs? Are these topics different from the ones in the newspapers? Finally, can unsupervised methods be used to help us understand these problems? Answering these questions is of paramount importance to understanding the relationship between foreign policy and mass media. Based on this discussion, this research has three main objectives: (a) to verify whether unsupervised methods can be used to analyze documents on international issues; (b) to understand the issues that politicians talk about when dealing with foreign affairs; and (c) to understand when and with which periodicity the mass media publish news on certain international topics. To do so, I created two new corpora, one with news articles published in the international section of two major Brazilian newspapers; and a corpus with all speeches made within the two Committees on Foreign Affairs of the National Congress of Brazil. I ran a topic model using Latent Dirichlet Allocation (LDA) in both. The results of this topic model show that LDA can be used to distinguish different international issues that appear in both political discourse and the mass media in Brazil. Additionally, I found that the LDA model can be used to identify when some topics are debated and for how long. The findings also demonstrate that Brazilian politicians and Brazilian newspapers are neither isolated nor unstable in what regards international issues.

Keywords: Political Speeches, Newspapers, Topic Analysis, LDA, Brazil

Resumo

Sobre o que os políticos falam quando discutem temas internacionais? Esses tópicos são diferentes daqueles que aparecem nos jornais? Finalmente, métodos não supervisionados podem ser usados para nos ajudar a entender esses problemas? Responder a essas perguntas é de suma importância para entender a relação entre política externa e mídia de massa. Com base nessa discussão, esta pesquisa tem três objetivos principais: (a) verificar se os métodos não supervisionados podem ser usados para analisar documentos sobre questões internacionais; (b) compreender sobre que assuntos os políticos falam quando lidam com relações exteriores; e (c) entender quando e por quanto tempo a mídia de massa publica notícias sobre determinados tópicos internacionais. Para tanto, eu criei dois novos corpora, um com notícias publicadas no caderno internacional de dois dos principais jornais brasileiros; e um corpus com todos os discursos feitos dentro das duas Comissões de Relações Exteriores do Congresso Brasileiro. Executei um modelo de tópico usando *Latent Dirichlet Allocation* (LDA) em ambos. Os resultados desse modelo de tópico mostram que ele pode ser usado para distinguir diferentes questões internacionais que aparecem tanto no discurso político como na mídia de massa no Brasil. Além disso, o modelo pode ser usado para identificar quando alguns tópicos são debatidos e por quanto tempo. Os resultados também demonstram que tanto os políticos como os jornais brasileiros não são isolados nem instáveis em relação a questões internacionais.

Palavras-chaves: Discursos Políticos, Jornais, Análise de Tópicos, LDA, Brasil

À minha família
Pelo amor e carinho

Contents

List of Figures

List of Tables

List of Abbreviation

BRICs – Brazil, Russia, India and China.

CAE – Comissão de Assuntos Econômicos

CAPADR – Comissão de Agricultura, Pecuária, Abastecimento e Desenvolvimento Rural

CAS – Comissão de Assuntos Sociais

CCJ – Comissão de Constituição, Justiça e Cidadania

CCJC – Comissão de Constituição e Justiça e de Cidadania

CCT – Comissão de Ciência, Tecnologia, Inovação, Comunicação e Informática

CCTCI – Comissão de Ciência e Tecnologia, Comunicação e Informática

CCULT – Comissão de Cultura

CDC – Comissão de Defesa do Consumidor

CDEICS – Comissão de Desenvolvimento Econômico, Indústria, Comércio e Serviços

CDH – Comissão de Direitos Humanos e Legislação Participativa

CDHM – Comissão de Direitos Humanos e Minorias

CDIR – Comissão Diretora do Senado Federal

CDR – Comissão de Desenvolvimento Regional e Turismo

CDU – Comissão de Desenvolvimento Urbano

CE – Comissão de Educação (Câmara dos Deputados)

CE – Comissão de Educação, Cultura e Esporte (Senado)

CESPO – Comissão do Esporte

CFFC – Comissão de Fiscalização Financeira e Controle

CFR – Committee on Foreign Relations'

CFT – Comissão de Finanças e Tributação

CI – Comissão de Serviços de Infraestrutura

CIDOSO – Comissão de Defesa dos Direitos da Pessoa Idosa

CINDRA – Comissão de Integração Nacional, Desenvolvimento Regional e da Amazônia

CLP – Comissão de Legislação Participativa

CMA – Comissão de Meio Ambiente

CMADS – Comissão de Meio Ambiente e Desenvolvimento Sustentável

CME – Comissão de Minas e Energia

CMULHER – Comissão de Defesa dos Direitos da Mulher

CPD – Comissão de Defesa dos Direitos das Pessoas com Deficiência

CRA – Comissão de Agricultura e Reforma Agrária

CRE – Comissão de Relações Exteriores e Defesa Nacional

CREDN – Comissão de Relações Exteriores e de Defesa Nacional

CSF – Comissão Senado do Futuro

CSPCCO – Comissão de Segurança Pública e Combate ao Crime Organizado

CSSF – Comissão de Seguridade Social e Família

CTASP – Comissão de Trabalho, de Administração e Serviço Público

CTFC – Comissão de Transparência, Governança, Fiscalização e Controle e Defesa do Consumidor

CTUR – Comissão de Turismo

CVT – Comissão de Viação e Transportes

dfm – document-feature matrix.

dtm – document-term matrix.

EDA – Exploratory Data Analysis.

EU – European Union.

FINEP – Financiadora de Estudos e Projetos.

H4N – Harvard-IV-4 TagNeg Dictionary.

http – Hypertext Transfer Protocol

https – Hypertext Transfer Protocol Secure

INFRAERO – Empresa Brasileira de Infraestrutura Aeroportuária.

LDA – Latent Dirichlet Allocation.

LIWC – Linquistic Inquiry and Word Count.

MERCOSUL – Southern Common Market.

NYT – The New York Times.

QTA – Quantitative Text Analysis.

RID – Regressive Imagery Dictionary.

SEC – U.S. Securities and Exchange Commission.

SSL/TLS – Secure Sockets Layer/Transport Layer Security

UN – United Nations.

UNCSD – United Nations Conference on Sustainable Development.

PDT – Partido Democrático Trabalhista (Democratic Labour Party).

PT – Partido dos Trabalhadores (Workers' Party).

PTC – Partido Trabalhista Cristão (Christian Labour Party).

WHO – World Health Organization.

Acknowledgements

Although this dissertation bears my name, it could only come into existence thanks to the help, both academic and emotional, of countless loved ones. In this space, I would like to thank some of these people, while being aware that I am not mentioning people who were also important.

I will start by thanking Professor Janina Onuki, who was been my thesis director since my first college degree and has given me academic, professional and personal support. Janina's dedication was never restricted to me or other students, but was rather directed to IRI-USP and the International Relations area in Brazil as a whole. This research is fruit of this institutional effort of hers. All the methodological knowledge I acquired at Caeni or as a visiting researcher at Princeton University, as well as the double degree program which this PhD is a part of, are a result of Janina's efforts.

This research, then, is lucky enough to have two houses: King's College London (KCL) and Universidade de São Paulo (USP). At King's, I would like to thank Professor Anthony Pereira for being my mentor and for the academic and professional advice he gave me. On the other hand, within IRI-USP and Caeni, I have always counted on the academic support of numerous professors and researchers, but I would like to thank especially Professor Amâncio and Umberto for their advice and incentives for me to continue learning quantitative methods. I would also like to thank all of IRI's staff members, who since my undergraduate years have helped me navigate the USP bureaucracy as efficiently as possible. A special thanks to Cris, not just for the coffee, but for cheering IRI up in a unique way.

One of the most important things that academic life gave me was the opportunity to meet incredible people. Besides the ones I mentioned above, there are many colleagues and friends that I had the opportunity to meet inside USP, KCL or the other academic institutions that received me. Some of these people are Andreia do Carmo, Fernanda Figueiredo, Flavio Pinheiro, Gabriela Ferreira, Galileu Kim, Ignacio Cardone, Leonardo Falabella, Marketa Jerabek, Murilo Zacareli, Pedro Feliú, Pietro Rodrigues and Rafael Magalhães.

There are still some people who, besides getting to know them inside these universities, I got to share my house with. Some of the special people that lived with me and taught

me important things about life and myself were: Lucas Maciel, the incredible kitchenette troupe (Erika Medina, Fernando Mouron, Lúcio Salles, Uirá Souto Melo and Francisco Urdinez), Lívia Prado, Isabela Mazão, Valentina Ferrari, Zoe Barossi, Camilla Cruz, Liselotte Willig and Sebastian Meiser. I would also like to thank the friends who have remained since the pre-college preparation course, or from long before that: Ana Cláudia Moreno, Priscila Kodato and Mirian Hayakawa.

Heartfelt thanks should also go to my whole family: aunts, uncles and cousins. I have the privilege of having a large family, who, despite its contradictions and differences, has always sought to be together. Special appreciation goes to two people: my cousin Cris and Graça, for having stuck with me since childhood. Cris, thank you for always listening and giving me affection and attention in my most difficult moments.

I thank Renata Mendes for having accompanied me throughout much of this journey, from statistics classes to the United States, and then in London. With affection, Renata always supported me and encouraged me to seek new challenges and to trust myself more. This attention was not limited by geographic or emotional distances. Thank you, Re.

Finally, I would like to thank my parents and brothers. Without their unrestricted support and affection, I would not have gotten here. Even with all the financial difficulties, my parents have always prioritized my studies and never spared efforts to help me finish my research projects. They were always very understanding and attentive to me, even when work prevented me from being physically with them. I also have to thank Pedro, my twin brother. Thank you for supporting me and for helping me at home during my absences. Finally, thanks to Raphael, my younger brother and the most incredible person I know. Despite all the physical limitations, the love and affection that Rapha transmits through his eyes are incomparable.

Rapha, thank you for making me a better person by teaching me to see the world with more joy and empathy!

1 Introduction

Which international issues do Brazilian politicians talk about? What is the role played by mass media when dealing with international news? Finally, is it possible to analyze political discourse and mass media in Brazil with an unsupervised method? Being able to understand the key issues of this debate is of paramount importance, for it would give us insights about how the characteristics of these actors interact to shape the way information on international issues affect public discourse on foreign affairs. As I will show throughout this dissertation, Quantitative Text Analysis can be used to analyze the way both politicians and the media in Brazil debate international issues.

The huge amount of textual information we produce nowadays means a challenge for the traditional analysis of content made by coders, given the elevated employment of financial and human resources that this method requires. These needs, however, can be mitigated through the use of quantitative text analysis methods, since they allow the analysis of extensive corpora, without prior human classification. That is why it is extremely important to evaluate unsupervised methods for quantitative text analysis.

An additional difficulty that I had to face in this thesis concerns a long debate on the nature of foreign policy studies in the area of International Relations. Is foreign policy different from other public policies, in terms of formulation, accountability, or the possibility of popular participation? This questioning is motivated by the disjunction between foreign issues and the daily life of ordinary citizens, which results in the low mobilization capacity that society presents when it comes to influencing the formulation of foreign policy (Almond, 1977; Lippmann, 2010).

According to some authors, such as Rosenau (1967), Rissen and Kappen (1995), and Holsti (2004), only issues that can directly impact people's lives –such as education, health-care and economy– are able to mobilize public opinion. Since an international issue has to be pivotal to have an impact on the public opinion, I intend to analyze political speeches in the National Congress of Brazil and news articles in Brazilian newspapers in order to identify which are the most salient foreign issues discussed in both sources.

Based on this discussion, this research has three main objectives: (a) to verify whether it is possible to use unsupervised methods to classify international issues debated in Brazil by politicians and by the media; (b) to understand which foreign issues are addressed in

Brazilian newspapers and with which periodicity; and (c) to understand which international topics the Brazilian politicians talk about, and when they do so.

In order to achieve these goals, I begin by gathering Brazilian newspapers articles and politicians' speeches regarding foreign issues. Therefore, I collected all the international news articles from two of the biggest Brazilian newspapers: *Folha de S. Paulo* (Folha) e *O Estado de S. Paulo* (Estadão). In terms of periodicity, I collected all the news published in the international affairs sections of Folha (*Mundo*) and Estadão (*Internacional*) between January 2000 and December 2018; much of the news, however, is concentrated between 2007 and 2017. The corpus of news on international issues has a total of 174,515 items, out of which 132,863 are Folha articles and 41,652 are Estadão's.

Moving on to the second corpus that I intended to analyze, namely, the one about politicians' speeches on foreign issues, I created a database with all the speeches made by politicians in both Committees on Foreign Affairs of Brazil's National Congress: *Comissão de Relações Exteriores e Defesa Nacional* (CRE) in the Senate, and *Comissão de Relações Exteriores e de Defesa Nacional* (CREDN) in the Chamber of Deputies. I have chosen these two Committees because they are specific to international relations, and because, since they have many members and guests, there is also a fairly large set of statements to be retrieved from them. In this sense, this research departs from the analysis of presidential speeches, which usually include only a few dozen speeches and a handful of presidents.

As an example of contrast to these works in terms of volume of analyzed data, this research analyzed speeches made within these two Committees on Foreign Affairs of the Brazilian Congress, CRE and CREDN, between January 2000 and December 2017, a period during which Brazil was governed by four different presidents: Fernando Henrique Cardoso (1995-2002); Luiz Inácio Lula da Silva (2003-2010); Dilma Rousseff (2011-2016); and Michel Temer (2016-2019). Meanwhile, for each legislative mandate, CRE is made up of 19 senators, and CREDN has around 36 deputies (this figure may vary with each legislature), not considering the guests (Congress members, scholars, diplomats, ministers, etc.) who also participate in the meetings of these two Committees.

Considering this large number of members and guests, the corpus of political speeches in the two Committees has a total of 62,410 oral statements held in the two Committees on Foreign Affairs between 2000 and 2017. These speeches are distributed among 44,005 statements in CRE (Senate) and 18,405 in CREDN (Chamber of Deputies).

After collecting the news and speeches, I structured these two corpora into databases. Then, I parsed the newspaper articles data and speech data so the final output would only contain the information associated with the research. Afterward, I ran a topic model using Latent Dirichlet Allocation (LDA) to identify 60 topics in the Committees' corpus and 80 topics in the newspapers' corpus.

The findings show that topic modeling can be used to analyze foreign affairs topics in Brazil. LDA creates coherent topics, and their connection to world events can be verified. As for the findings of the topic model for newspaper items, the LDA model shows that Brazilian journalism follows the line of war journalism rather than that of peace journalism. Moreover, results show that politicians are not isolated from the debate on international events, as they often discuss issues that are affecting Brazil or the world. Nevertheless, the relationship between mass media and political discourse could not be proven.

As for the structure of the thesis, except for the Introduction and the Conclusion, each chapter was thought of as an independent academic article. In this sense, it is important to point out to the reader that there is a similar technical information in each chapter, as they were intended to be published separately. Another consequence of this is that the reader does not necessarily need to read the chapters in their order of appearance.

The thesis is structured divided as follows: in Chapter 2, I make a bibliographical survey on the use of quantitative methods of classification, by debating the use of three types of models: supervised, dictionary-based and unsupervised method. Then, in Chapter 3, I describe the process of collecting data from newspapers about international affairs articles, as well as the results of the topic model for this first corpus. In chapter 4, besides presenting the structure of the two Committees on Foreign Affairs of the Brazilian Congress, I explain the step-by-step process for collecting speeches made in these Committees and discuss the results found by the topic model for this second corpus. Finally, in chapter 5, I present the conclusions raised by this research.

2 Treating Text as Data

Text analysis is a powerful tool to understand political discourse. Nonetheless, the amount of information that is produced nowadays makes it virtually impossible for a researcher to analyze it all manually. Moreover, the analysis of such information faces constraints not only due to quantity, but also in terms of budget. Before the evolution of computational processing, only well-financed projects were able to hire human coders to analyze vast amounts of documents.

This scenario has changed along with technological evolution. Today, it is possible to analyze a great amount of texts with personal computers and at a low cost. At present, the vast majority of computers on the market can handle the storage and processing of data from thousands of documents in a few minutes. Another key point that helped raising this new paradigm was the creation of free software for quantitative data analysis, such as the *quanteda* and *tm* packages in *R* programming language.

In this thesis, I'll take a step-by-step look at how we can treat text as data, inspired by the work of Grimmer and Stewart (2013). More specifically, I'll describe how can we find and assign a set of documents into categories, previously defined or not. For this, one of the first steps is getting to know in depth the set of texts to be analyzed.

This corpus, should contain only texts with the same theme or subject, since, in terms of efficiency, QTA works best when dealing with only a specific issue. If part of the documents is not relevant to the analysis, then the researcher must cut them out. For example, if the research objective is to classify political discourse, the corpus should bear only texts of political speeches, because all other information will eventually impact the results.

Second, the length of the documents being analyzed is another important characteristic for an efficient QTA. The longer the text, the better the model. Given that QTA models are based in how words are used, how often the words appear and how often they appear together, longer texts provide more examples for the model. Finally, if the research intends to analyze shorter texts, a necessary precaution is to analyze a huge quantity (thousands or even millions) of short texts.

After this careful analysis of the corpora, we will be able to apply some QTA techniques. One of the most common and useful techniques is classification. This algorithm enables

the assignment of a large amount of text to "boxes" or categories, both predefined or not. In this chapter, I will discuss the different techniques of textual classification: supervised, dictionary-based or unsupervised.

In terms of structure, first I will describe the step-by-step for preparing the set of texts to be analyzed quantitatively. Next, I will describe the types of existing classification algorithms, their strengths and limitations. Finally, in the conclusion, I will sum up the main ideas discussed in this section.

2.1 Preparing documents to be analyzed

In this thesis, before deepening the analysis of the two corpora –political speeches and newspaper news–, I will first discuss the methodological steps that must be taken. I will also discuss the main existing quantitative text analysis methods and what their strengths and weaknesses are.

With this objective, and after checking the characteristics of the corpus as discussed in the last section, the researcher has to transform text into data, a process in which information is invariably lost due to the complexity of language. However, that does not mean that QTA cannot be used to analyze discourse; it only means that the researcher should be aware of it, and should be careful in selecting which information to retain and which information to discard for not being useful to the analysis (Grimmer and Stewart, 2013, p.272).

An important distinction that has to be made is that building quantitative text analysis methods is different from causal model building. In the latter, one has to include all the variables that affect the data-generation process. This principle, however, does not apply to the quantitative text analysis model, for in QTA including more features does not guarantee a better model, nor does reducing the number of variables being analyzed mean a worst model. Therefore, we can select and reduce the amount of information that will serve as input for a quantitative text model knowing that we are actually creating a better model.

One of the first steps in QTA that results in loss of information is treating documents as a *bag of words*, which means that the order in each words appear does not matter for the analysis. Although it seems to be a fairly unreasonable assumption, since the order of words in a sentence can change its meaning, in practice, such phenomenon is rare. In

general, texts or speeches are built with an idea that does not take the order of words into account. Even more important, the methodological processes, such as topic modeling and sentiment analysis are not improved when considering the order in which the words appear (Manning et al., 2008; Hopkins and King, 2010a).

Despite that, if the researcher finds that word order is important, the n-gram can take order into consideration, when n is equal to or greater than two. When we treat documents as *bag of words*, n is equal to one, and we call it a unigram, since documents become a "simple list of words" (Grimmer and Stewart, 2013, p.272). When n is equal to two, the text analysis is done by pairs of words, which is known as bigram. In bigram analysis, terms such as "human rights" or "public policy" can be interpreted: instead of dealing with a simple "list of words," bigram matrices compare dyads.

Following the same pattern, trigram is a comparison of three words. In this case, terms like "foreign affairs committee" and "war on terror" will appear as a token –a token is an unit of analysis in a quantitative text analysis. Given that most QTAs treat documents as *bag of words* or unigrams, a token is typically a word, a punctuation mark or a number. When using bigrams, the token is a pair of these elements (word and word *or* word and number, and so on). Therefore, even though the simple version of QTA ignores word order, there are some steps that the researcher may take in order to mitigate this issue.

Besides removing word order, another step to reduce complexity in text analysis is removing punctuation marks and numbers from the corpora. The case for discarding them is straightforward: they do not provide any useful information for the text analysis. Counterintuitively, though, we also remove very common words, that is, those appearing in more than 99% of the documents, and rare words, which occur in less than 1% of the documents.

In the first case, we remove them because they appear so often they do not provide any information that could help differentiate the documents. In the other case, rare words will not provide sufficient cases for the algorithm to pinpoint the difference between documents; that is why we discard them. Finally, we filtered out words known as *stop words*, such as *the*, *is*, and *it*, because they belong to the very common words scenario. There is not a universal list of stop words, but most packages in *R* for QTA (e.g. *quanteda* and *tm*) have a built-in list that the researcher can use to remove stop words from documents.[1]

[1]The *quanteda* package has a list of stop words for several different languages (including Portuguese), which is known as *Multilingual Stopword List*, and can be accessed at: `http://stopwords.quanteda.io/`

Moreover, besides disregarding word order and filtering out words that do not provide useful information, we have to transform all capital letters into lower cases, given that the algorithm distinguishes between the two. Therefore, "House" and "house" are two different tokens, even if they carry the same meaning. This process is quite simple in computational terms, since the computer only has to change uppercase letters by lowercase ones.[2]

All the processes described above are designed to reduce the number of tokens in the corpora, which lowers the complexity of the model, making it smaller in size so the computer can calculate it faster. However, there is one more step that we can take to reduce the number of tokens without losing important information. We can use a stemming algorithm or lemmatize the words; the former only considers word by word, while the latter depends on the context in which the word is being used.

A stemming algorithm reduces the word to its stem or root form by removing prefixes and suffixes. For example, "unlike," "likely," "unlikely," "likes," "liking," and "liked" become "like." Nonetheless, there are different types of stemming algorithm that can be used to reduce corpora's size. One of the most common types of rule-based stemming algorithms is known as Porter (Porter, 1980). The popularity of the Porter algorithm arises from the simple way it finds the stem of the word. Sometimes, however, the word produced by the Porter stemming algorithm is not an actual word, as "ties," for instance, becomes "ti."

Although Porter's stemming algorithm is fast and most of the times produces intelligible outputs, the results sometimes are barely recognizable; in fact, unintelligible results is the biggest downside to it.[3] Nonetheless, if the researcher wants to reduce dimensionality with a stemming algorithm while maintaining the output as real words, they can use the Krovetz stemmer. Because it is dictionary-based, the Krovetz stemmer produces real words as an output. The disadvantage of Krovetz's algorithm is that it depends on a dictionary, which means that each language used requires a different dictionary.

On the other hand, lemmatization does a morphological analysis of the words, instead of having a list of suffixes and prefixes to remove as a stemmer algorithm does. Therefore, rather than a dictionary-based analysis, lemmatization is a linguistic methodology, because it considers the meaning of the word in a sentence. Hence, a lemma can change depending

[2]The main command of *R base* (*R*'s basic library) that does this task is `tolower`, but it is also possible to execute this step with the package *quanteda* when transforming words into tokens or creating the document-feature matrix, respectively, as follows: `tokens_tolower()` and `dfm_tolower()`.

[3]There are other types of rule-based stemming: *Lovins Stemmer* is a list of suffixes to be removed; *Dawson Stemmer* is an extension of the *Lovins Stemmer*.

on which context the word is being used. For example, the lemma of "better" is "good," while its stem is also "better." In spite of that, there are lemmas that are equal to their stems: both the lemma and the stem of "walking", for instance, are "walk."

The advantage of using lemmatization au lieu de a stemming algorithm is that lemmas are always real words, while stems sometimes are not, such as "*famili*" from "*families*." This aspect makes the process of interpreting topic models more direct and intelligible when using lemmas. Notwithstanding this positive aspect, there are two main inconveniences to using lemmatization: first, it requires a dictionary, and not all languages have a lemmatizing dictionary available, at least not with free public access;[4] second, given that the lemmatization algorithm takes context into account, the process takes more time and needs more processing power as compared to a stemming algorithm.

As explained, all the steps mentioned above are designed to reduce the amount of tokens in the analysis without losing important information. The output of this process is what is known as *document-term matrix*, where the documents are in rows and each term or token is in a column. By discarding these "uninformative" tokens, the amount of columns is reduced and the *document-term matrix* gets smaller. The cells of this matrix register the number of times that each token appears in each document. Therefore, in the *document-term matrix* we can know not only if the token was used in that document, but also the number of times that each token occurred in each document.

Table 1: An example of a *document-term matrix*

	congress	land	project	...
Document_1	0	0	1	
Document_2	2	0	3	
Document_3	1	0	2	
...				

Finally, the researcher has to pay attention to the amount of *features* (terms or tokens) and to the amount of zeros that the *document-term matrix* has, which is known as *sparsity*. Usually, a typical *document-term matrix* has between two thousand and five thousand unique features. This number will depend on the type of vocabulary used in the corpora

[4]Luckily for this research, there is a lemma dictionary for Portuguese. I applied this dictionary to the two corpora analyzed herein: political speeches and newspaper news. Although this dictionary produces small errors, like turning "fronteira" [noun, border] into "fronteirar" [verb, put something in front of another], these errors were rare and did not affect the interpretation of the topic models performed in both corpora.

and the length of each document. *Sparsity* is a measure of how "empty" (cells with zero) the matrix is, and it varies between 0 and 1. The closer sparsity gets to 1, the more zeros the matrix has. A typical *document-term matrix* has an average sparsity between 0.8 and 0.9. If a column has sparse equal to 1, that means that the feature do not occur in any document and will be removed from the analysis.

Steps to transform documents into quantitative data:

1. Creating a bag of words (the order in which words appear is not relevant) = Unigram – because more complex forms do not increase performance in the analysis (Manning et al., 2008; Hopkins and King, 2010a), but it is also possible to do bigrams or trigrams (Martin and Jurafsky, 2009).

2. Stemming: reducing the word to its root and, by doing that, reducing the "dimensionality" of the text (Grimmer and Stewart, 2013, p.6). The most common algorithm to stem is known as Porter (Porter, 1980). Another method that can be used is lemmatization, which consists in reducing the word to its base form (better = good) using dictionaries and context analysis. The former method, stemming, is more simple and faster, when compared with lemmatization.

3. Removing punctuation marks, stop words, uncommon or very common words (usually those that appear in less than 1% of the documents and those that appear in more than 99% of the documents); and transform all words into lower case (one has to to be careful with acronyms).

4. Output: Like Benoit et al. (2018), I use the term "document feature matrix" (dfm), but it is also known as "document-term matrix" or “term-document matrix." Regardless of the nomenclature, this matrix bears each document in a row, and in the columns there is every unique word (token or feature) used in the corpus. The cells contain the frequency with which each features occur within each document.

These are general instructions, but the steps that each researcher will have to take depend on their research question and type of data used. For example, if the objective is to carry out a grammatical gender analysis (Monroe et al., 2008), it is a bad idea to delete gendered pronouns. Another important aspect that the researcher will have to decide is

whether the order of appearance of words is relevant or not, for sometimes the order is meaningful and has to be included in the analysis.

Given that different methods of QTA can be used to analyze an enormous collection of documents, it is the researcher's responsibility to verify if the results are coherent. Therefore, it is the interest of the researcher to show proof demonstrating that their results are valid and reasonable. Grimmer and Stewart (2013) emphasized the difference between validating a result from an unsupervised method and the output of a supervised method.

In both cases, researchers have to demonstrate that the results are a replication of what humans would do. However, given that in the unsupervised method we do not have pre-established human categories, the validation process of it needs a combination of different processes, such as statistical and substitutive evidence that leads to the conclusion that the unsupervised results are similar to a hypothetical supervised method trying to accomplish the same task.

Additionally, the validation process has a step of internal validity and one of external validation. The internal validity phase is responsible for verifying that the data found by the classification algorithms is coherent. Since the assignment of documents into "boxes" or into a spectrum follows quantitative parameters, results obtained may not be coherent. An example of low internal validity would be the assignment of very distinct texts into the same category. Therefore, in order to analyze and determine if the results have internal validity, the researcher has to resort to their knowledge on the subject being studied.

Once the researcher has assessed the internal validity, the next step is to analyze the external validity of results. At this stage, the researcher should use the literature to compare the results obtained with results from previous studies to check for consistency. This external validation, on the one hand, can be methodological, when the comparison is based on the method and verifies if the method used in the research had results similar to the methods applied in other sets of texts.

On the other hand, this external validation can be substantive, when the literature is used to verify whether the results obtained are theoretically coherent. This substantive analysis is especially important when the quantitative analysis of text is done in a set of texts never previously analyzed. In this situation, as the set of texts is new, the methodological comparison is not possible, since there are no previous comparable studies.

It is possible, however, to validate the results in a substantive way, for example, by pointing out that the literature indicates that such polarization tends to occur in situations similar to those found in the text set.

These validations must be made regardless of the type of method used, or even the source from which the texts were collected. This is because, nowadays, it is possible to extract texts from different sources by web scraping (Jackman, 2006); by OCR (Eggers and Hainmueller, 2009); or, in the most difficult cases, such as captcha blocking the automatic web scraping, we can use Mechanical Turk (Berinsky et al., 2012) to obtain these documents. In spite of the profusion of new methods to acquire documents, some documents are more suitable for quantitative text analysis. Grimmer and Stewart (2013) point out three aspects that make a document more advisable for use in QTA: (1) the text focuses the topic that is being considered in the analysis (classification method); (2) the text expresses political positions that are being scaled (scaling method)[5]; and, finally, regardless of the method, (3) the document has to have a certain length, because most QTA methods rely on a sufficient amount of words to work.

After performing all these steps, the *document-term matrix* is ready to be used. In the following sections, I will discuss one main document analysis method in which quantitative text analysis can be used: classification into categories. This process can be done either in a supervised way –when the researcher impute some conditions or pre-categorized documents into the model–, or in an unsupervised way–when the algorithm does not have any examples to follow or to use for testing.

2.2 Classification into categories

Classification into categories is one of the possibilities of quantitative text analysis. In fact, separating a set of texts into categories is one of the tools most used by political science as regards quantitative text analysis algorithms.

There are three ways of doing this classification: 1) supervised, such as by using a set of pre-classified documents; 2) unsupervised, where the mathematical model classifies the set of texts based on characteristics of the next documents; 3) dictionary-based classification, where the classification model tries to find a set of preselected words, a dictionary, in the

[5]In this work, I will not discuss the scaling method, since my purpose is to classify texts into categories. However, there is a whole area of quantitative text analysis dedicated to assigning texts in a continuum (scaling method), instead of putting them into "boxes" (classification method).

set of documents. In this section, I will discuss each of them.

2.2.1 Supervised method

Among the different ways of categorizing a set of texts, the supervised method stands out for the easy validation of its results and for the conceptual care that it obliges researchers to have (Hillard et al., 2008; Stewart and Zhukov, 2009; Hopkins and King, 2010b). This happens because this method requires a set of documents previously encoded by humans, known as training set, which will serve to train the algorithm. After the algorithm is optimized, it is applied to a set of new texts to be categorized. Finally, the result of the classification of this new set is validated.

These three steps (training set, algorithm application, validation) must be present in all supervised classification models. Given these characteristics, the supervised classification model is widely used in human sciences. This is because, as we will see later, the validation process is easier in the supervised method than in the dictionary-based method. In addition, since the supervised method depends on the existence of a set of pre-coded texts, this type of model requires the researcher, from the beginning of the research, to elaborate precise and mutually exclusive concepts that will base the classification process.

The work of Stewart and Zhukov (2009) gives us an example of how this method can be used. In one of the stages of the paper, the authors used a supervised method to analyze the public debate about the use of force in Russia. In order to do so, the researchers collected 7,920 public statements made by politicians or by the military high command between 1998 and 2008, a period that corresponds to the Georgian-Ossetian Conflict that culminated in 2008 in the Russo-Georgian War.

This is then a key period in Russian history, where researchers can see how the military and politicians address the issue of Russian involvement in conflicts and the use of the Russian military in international conflicts. The use of the supervised method in the analysis of these 7,920 public statements was possible because a test database was created with 300 documents. These 300 documents were randomly selected from the 7,920-document pool and were manually coded.

With this manual coding, the supervised method allows for a predetermined categorization. As opposed to the unsupervised method –in which there is no *a priori* control over which categories the algorithm should use as a base for assigning the documents–,

the supervised method allows a control on not only the number of categories, but also on the substantive characteristics of each category. Therefore, the supervised method is indicated when there is a need to assign texts into pre-delimited categories.[6]

Stewart and Zhukov (2009) have manually coded these 300 documents into two categories, "Activist" or "Conservative." In the first case, the statement was considered "Activist" if: (a) it supported the use of force; (b) if it was in favor of unilateral solutions for the resolution of international conflicts; (c) or if it had a revisionist stance of the international system as opposed to maintaining the status quo. On the other hand, the texts were deemed as "Conservative" when they: (a) affirmed that the use of force should only be employed as a last resource; (b) expressed doubts about the possibility of solving international conflicts through the use of force; (c) encouraged/recognized the need for multilateral action to resolve international conflicts.

The procedure described above is the first step of the supervised method. Creating pre-established categories is fundamental, because this set of documents, called test set s, will be used as a training base for the model. The second part of the supervised method, which is the choice and refinement of the model, was made as follows by Stewart and Zhukov (2009): (a) randomly selecting 275 documents from the 300 training base documents; (b) training the model based on these 275 documents and then testing the remaining 25 documents of the training base; (c) simulating the refinement process 10,000 times. In the end, the authors had an algorithm accurate enough to apply to the 7,620 documents, which were not classified.

The results of Stewart and Zhukov (2009) found only a quarter of the 7,920 public statements to be "activist." The great majority of public statements made by Russian politicians and military were "conservative" between 1998 and 2008. Although there were far fewer "activist" statements, Stewart and Zhukov's (2009) results show that members of the Russian Armed Forces have a more activist stance as compared to politicians, who proved to be largely conservative in terms of the use of force in international conflicts.

Finally, the third step of the supervised method, validation, was also carried out in Stewart and Zhukov's investigation (2009). In this phase, the authors point out that the findings of the supervised model were consistent with the literature of Russian political science, since several authors maintain that the Russian military believe in the use of

[6] In contrast, the unsupervised method is more exploratory, since it does not require categories to be defined before starting the analysis.

force as an international conflict resolution instrument. Thus, the text by Stewart and Zhukov (2009) is a good guide on how supervised models of quantitative text analysis are employed. Next, we'll see how we can do this categorization process by using a dictionary.

2.2.2 Dictionary-based method

In the previous section, I analyzed the use of supervised methods to categorize a set of texts, a process in which it is essential to have a set of pre-classified texts to serve as a training base for the classification algorithm. In many cases, however, there is no set of pre-coded texts, or the intention of the research is to classify in general terms whether the texts are positive or negative.

In these situations, the text classification method can use dictionaries. Instead of counting how many words are being used in each text, as in the supervised method, the dictionary method counts the occurrence of predefined words (*values*), which are associated with canonical concepts or terms known as *keys*. Therefore, the dictionary associates several values with their respective keys.

Depending on the dictionary, this association is exclusive, that is, if a value belongs to a key, it will not be associated with any other key. An example of this is the dictionary that contrasts positive values with negative ones, as the General Inquirer Dictionary does. In this dictionary, the word "admire" is in the positive key and therefore will not be associated with the negative key. However, there are dictionaries where a value can be associated with more than one key. This occurs in dictionaries where different aspects of the meaning of a value may correspond to different keys. For example, the word "cried" is associated with five different categories (past tense verb, verb, overall affect, sadness, negative emotion) in the Linguistic Inquiry and Word Count Dictionary (Pennebaker et al., 2007).

This differentiation between multiple association or not leads us to discuss one of the main questions regarding the use of dictionaries in the quantitative analysis of text. Namely, the researcher must know in depth the objectives and characteristics of the dictionary to be employed in the analysis. There are several dictionaries available on the internet, both for free (General Inquirer Dictionary) and paid (Linguistic Inquiry and Word Count). It is important to note that the fact that there is a company behind the dictionary does not mean that it is better than a free one. This is because dictionaries are built for specific purposes, and the researcher must choose the one that best suits the

objectives of a specific query.

Next, I will describe three of the main dictionaries, and the needs they were developed to meet. The first is the General Inquirer (Stone et al., 1966), which serves to classify whether a text has positive or negative connotations. This dictionary has a total of 4,206 words distributed in two categories: positive (1,915 words) and negative (2,291 words).[7]

Table 2 illustrates some of the value entries in the General Inquirer Dictionary. On the left side of the table we have the words associated with a positive key. In turn, on the right side, I listed some examples of words belonging to a negative key.

Table 2: General Inquirer Dictionary example entries

Positive	Negative
able	abandon
abound	abnormal
accept	abrupt
acclaim	absurd
accord	addict
accuracy	adverse
achieve	afflict
adequate	against
admire	anarchy
affirm	antitrust
...	...

The second dictionary I am going to present is more complex in structural terms than the General Inquirer Dictionary. It is called Regressive Imagery Dictionary and was created following the ideas developed by Martindale, Colin (Martindale, 1975; Martindale, Colin, 1990). The English version of this dictionary has 3,200 words assigned into 43 categories, distributed in three sets of categories: 29 categories of primary cognitive processes; 7 categories of secondary cognitive processes; and 7 categories of emotion.[8] The main as-

[7] In 2000, the General Inquirer Dictionary was updated and gained many other categories. The new version of this dictionary has 182 categories, far beyond the initial two, and was renamed "Harvard IV-4." The new version can be accessed at: `http://www.wjh.harvard.edu/~inquirer/Spreadsheet.html` – Last visited on April 27, 2019.

[8] There is a version of the Regressive Imagery Dictionary in Portuguese, which was translated by Tito Cardoso e Cunha, Brigitte Detry and Robert Hogenraad. The Portuguese version can be accessed at: `http://www.provalisresearch.com/Download/PRID.ZIP` – Last visited on April 24 2019.

sumption of this dictionary is that psychological processes can be observed through text, and that there is a division between primordial and abstract processes:

> The Regressive Imagery Dictionary (Martindale, 1975,1990) is a content analysis coding scheme designed to measure primordial vs. conceptual thinking. Conceptual thought is abstract, logical, reality oriented, and aimed at problem solving. Primordial thought is associative, concrete, and takes little account of reality.
>
> (Kovach Computing Services – Wordstat – 2019)
>
> Link: `https://www.kovcomp.co.uk/wordstat/RID.html`

Table 3 lists some sample words in the "Sample" column and shows which categories they are associated with in the Regressive Imagery Dictionary (RID).[9] In comparative terms, as mentioned above, this dictionary is structurally much more complex than the General Inquirer (first version).

This complexity is due to the very purpose of the dictionary. While General Inquirer initially intended to analyze whether the text was positive or negative, RID aims to understand both the logical aspects (conceptual thought) and the most unrealistic or imaginative features of the text (primordial thought). Thus, the researcher who will use a dictionary to do quantitative text analysis should be well aware of the goals for which the dictionary was built and whether they match the research objectives.

[9]For presentation reasons, I have not listed all sub-subcategories of the subcategories "Sensation," "Defensive Symbolization," "Regressive Cognition," and "Icarian Imagery.". That is why there are some sub-subcategories represented by "...".

Table 3: Regressive Imagery Dictionary example entries

Category	Subcategory	Sub-subcategory	Sample
Primary Process	Drive	Oral	Breast, drink, lip
		Anal	Sweat, rot, dirty
		Sex	Lover, kiss, naked
	Sensation	General Sensation	Fair, charm, beauty
		Touch	Touch, thick, stroke
		Taste	Sweet, taste, bitter
		...	...
	Defensive Symbolization	Passivity	Die, lie, bed
		Voyage	Wander, desert, beyond
		Random Movement	Wave, roll, spread
		...	...
	Regressive Cognition	Unknown	Secret, strange, unknown
		Timelessness	Eternal, forever, immortal
		Consciousness Alteration	Dream, sleep, wake
		...	...
	Icarian Imagery	Ascend	Rise, fly, throw
		Height	Up, sky, high
		Descend	Fall, drop, sink
		...	...
Secondary Process	Abstraction		Know, may, thought
	Social Behavior		Say, tell, call
	Instrumental Behavior		Make, find, work
	Restraint		Must, stop, bind
	Order		Simple, measure, array
	Temporal References		When, now, then
	Moral Imperative		Should, right, virtue
Emotions	Positive Affect		Cheerful, enjoy, fun
	Anxiety		Afraid, fear, phobic
	Sadness		Depression, dissatisfied, lonely
	Affection		Affectionate, marriage, sweetheart
	Aggression		Angry, harsh, sarcasm
	Expressive Behavior		Art, dance, sing
	Glory		Admirable, hero, royal

Finally, there is the Linguistic Inquiry and Word Count (LIWC) dictionary, which, in

terms of its hierarchical structuring, somewhat follows the RID line. LIWC was developed by Pennebaker et al. (2001) "for studying the various emotional, cognitive, and structural components present in individuals' verbal and written speech samples" (Pennebaker et al., 2007, p.3). Later, the LIWC was reviewed by Tausczik and Pennebaker (2010) and gained new categories. Currently, it has around 4,500 words and word stems, assigned to one or more of the 82 dimensions of language.

One of the upsides of LIWC is that it has been translated into several languages: German, Dutch, Italian, Portuguese, Norwegian and Spanish.

The description of the objectives and structures of these three dictionaries allows us to observe that the quantitative analysis of text becomes quite rich with the use of a dictionary. However, this method has some points that the researcher must be aware of when considering applying it to a text set. As I mentioned above, the purposes or assumptions of a dictionary should be consistent with the research to be developed. Namely, if the goal is to identify positive and negative texts in a set of documents, one can use the General Inquirer Dictionary, but not RID or even LIWC.

When there is no specific dictionary for the dimensions to be analyzed, the correct step is for the researcher to create their own dictionary. The article by Laver and Garry (2000) describes the process of analyzing party manifestos with a dictionary created by them. This was the appropriated solution, since their purpose was to understand the political position of political parties in England and Ireland, and there was no dictionary elaborated with the aim of verifying political positions through text.

Thus, Laver and Garry (2000) apply manual coding methods and dictionary-based quantitative text analysis to verify whether the results of the computational method were valid internally and externally. Finally, the authors validated the results of the computer estimation with the results of expert surveys, but also with previous analyses of party programs.

The findings of this work indicate that the use of a dictionary in the quantitative analysis of text to study party manifestos was valid both internally and externally. The results obtained by Laver and Garry (2000) show a "high degree of cross validation" (Laver and Garry, 2000, p.619) between manual and automatic coding.

Thus, the dictionaries applied to the dimension for which they were constructed bring valid results. However, in addition to the dimension, the researcher must observe if the

theme of the set of texts to be analyzed follows the rules of the dictionary to be used. This question was raised in the article by Loughran and Mcdonald (2011), in which the authors analyze the negative feeling of 50,115 firm-year 10-K filings[10] between 1994 and 2008. For that purpose, the authors used the Harvard-IV-4 TagNeg(H4N), part of the Harvard Psychosociological Dictionary.

As this dictionary was designed to capture negative aspects in the texts, one could imagine that by applying H4N to the corpus bearing the 10-K reports, the authors could verify the negative market sentiment. However, the results obtained by Loughran and Mcdonald (2011) show that the use of this dictionary for the corpus in question is not valid. Therefore, although the research question and the dictionary used are in the same dimension (negative field), the dictionary did not produce valid results. This occurred because H4N was applied to a thematic area –the financial context– in which words that usually have negative connotations are not negative.

The authors show that almost 75% of the negative words in H4N had no negative connotation in the corpus of 10-K reports. Words classified as negative in the H4N, such as "mine," "cost," "board," "tax," "foreign," "capital," are not negative in the financial context. In addition, the authors show that the opposite fact also occurs: words that were negative in the financial context (e.g. "felony," "unanticipated" and "litigation") were not classified as negative in the H4N dictionary. Therefore, the researcher must be aware of the size and thematic area of his corpus and verify if the dictionary to be used in the quantitative analysis of text is valid. In the following section, I will describe the method of unsupervised quantitative text analysis.

2.2.3 Unsupervised method

Unlike the supervised method and the dictionary-based method explained in the previous sections, the unsupervised method does not require a set of pre-coded documents or a dictionary. Therefore, as we will see later, the unsupervised method, because it does not impose any predefined categorical structure, is extremely useful for exploring a set of texts that has not yet been explored by the literature, while producing theoretically interesting

[10] 10-K is a report required by the U.S. Securities and Exchange Commission (SEC) that bears data on the company's financial performance. Any company with operations in the United States that is worth more than $10 million and has over 2,000 owners (holders of equity securities) must publish the 10-K report annually. In addition to the financial data, this report details the company's history, its corporate structure and other data.

results.

Regarding the types of models, Grimmer and Stewart (2013) make a distinction within unsupervised methods between *Fully Automated Clustering* (FAC) and *Computer Assisted Clustering* (CAC). FAC is used to estimate the number of categories in the set of documents so that, based on this number, the algorithm can separate the documents into these categories (Jain et al., 1999; Manning et al., 2008). The CAC model, on the other hand, instead of applying only one clustering model as FAC does, tests several categorization models by cluster (Grimmer and King, 2011).

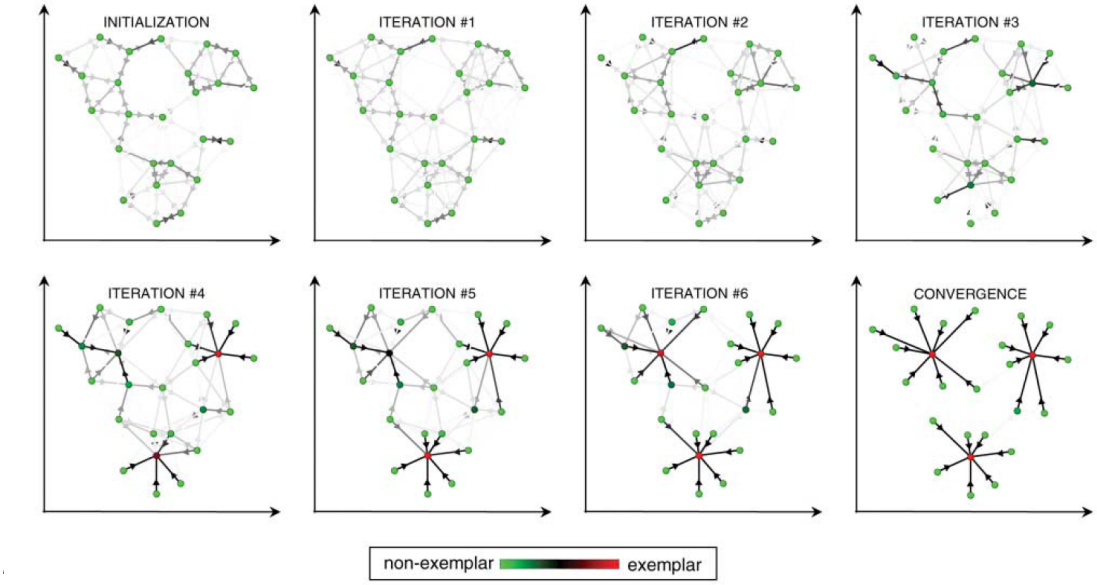
Grimmer and Stewart (2013) further point out that there are two ways to run a FAC: single membership models and mixed membership models. The central difference between these two models is whether the categories to which documents are assigned are mutually exclusive or not. That is, if each document is assigned to only one category (single membership model), or if it can belong to more than one category (mixed membership model).

The choice between these two methods must be based on theoretical and substantive terms. Therefore, before running a FAC model, the researcher should ask the following questions: does each document contain only one topic or express only one idea? Or can each document contain more than one topic or be classified in more than one category?

Depending on what one expects to find, the researcher can assume that the set of texts has a logic of single membership model, and run a model of K-means. The classification by the algorithm of K-means employs an optimization method for assigning texts to different clusters, and documents are assigned to the cluster that has the nearest cluster center (centroid).

Figure 1, elaborated by Frey and Dueck (2007, p. 973), shows how the K-means algorithm assigns documents to each interaction. In general terms, the K-means algorithm, at each interaction, attempts to minimize intra-cluster variation and maximize variation between different clusters. Interactions stop at the moment when the centroid change does not optimize the result, nor does the reassignment of observations into the different clusters.

Figure 1: How the K-means interaction works



Source: (Frey and Dueck, 2007, p.973)

Another point that is quite explicit in Figure 1 is the central feature of the single membership model that K-means algorithms have. This occurs because, as we can see, at the end of the interactions each point is connected to only one centroid, and there are no cases of assignment to more than one cluster/category.

However, there may be cases where the same document or speech addresses more than one theme or category. In this scenario, we need to use the mixed membership models, since these models assume that assignment to more than one cluster is possible. Among the several existing mixed membership models, one of the most used is the *Topic model* (Blei et al., 2003).

According to Grimmer and Stewart (2013), the Topic model has two central characteristics. The first is about defining the topic.

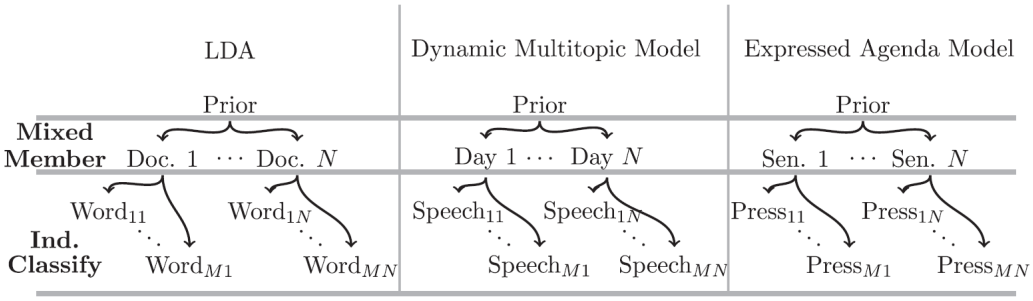
> Statistically, a *topic* is a probability mass function over words. For a topic k ($k = 1, \ldots,$ K) we represent this probability distribution over words with an $Mx1$ vector θ_k where θ_{mk} describes the probability the *k-th* topic uses the *m-th* word. Substantively, topics are distinct concepts. In congressional speech, one topic may convey attention to America's involvement in Afghanistan, with a high probability attached to words like `troop`, `war`, `taliban`, and `Afghanistan`. A second topic may discuss the health-care debate, regularly using words like `health`, `care`, `reform`, and `insurance`. To estimate a topic,

the models use the co-occurrence of words across documents.

(Grimmer and Stewart, 2013, p.17)

The second is about the hierarchical aspect with three levels that the topic model has. Figure 2, by Grimmer and Stewart (2013, p. 18), illustrates and exemplifies the difference between these three levels. While the first level is about the primary assumptions of the model, the second level holds the observations that can belong to multiple categories. Finally, the third level contains the items that impact the formation of categories.

Figure 2: Mixed membership model examples



Source: (Grimmer and Stewart, 2013, p.18)

In order to demonstrate how these different categories can be related, in Figure 2 Grimmer and Stewart (2013) listed three types of topic model and their respective hierarchical structures: (a) documents are an amalgam of all themes, since each word impacts each topic differently –Latent Dirichlet Allocation (LDA) (Blei et al., 2003); (b) each day has a set of themes that are impacted by different discourses –Dynamic Multitopic Model (Quinn et al., 2010); or senators' attention is divided among different press releases –Expressed Agenda Model (Grimmer, 2010).

Another aspect of the topic model that can be extracted from Figure 2 is the versatility that this type of model allows in contrast to the single membership models. Also, this type of model allows for customization in terms of temporal ordering or per document author –e.g. the order in which speeches are pronounced may affect the topic modeling.

In short, as we have seen in this section, there are several ways of classifying a set of texts. However, regardless of the unsupervised method used, the validation step is fundamental. This is because we must verify whether the categories produced by the algorithms are consistent with each other –internal validation–, and with what is discussed

by the literature –external validity–, for cases where there is empirical or substantive literature on the subject.

This validation step is even more important when we use unsupervised methods, since we do not have an *a priori* reference about the characteristic that the classification should have. Therefore, the researcher should have an extra concern when validating the results of unsupervised classifications. In the next section, I will discuss the scenario in which the researcher's objective is to arrange a set of texts in a specific continuous spectrum.

2.3 Conclusion

Computational evolution enabled the exploration of a vast amount of textual information. Researches that previously required the hiring of human coders and therefore substantial financial resources can nowadays be done with a personal computer. Despite this possibility, treating text as data through computational algorithms does not replace the work of the human encoder, but rather complements it.

Besides, before beginning to treat text as data, the researcher must take some steps. First, all texts to be analyzed must approach the same theme. For example, if the purpose of the research is to analyze presidential speeches, all documents in the corpora should be from presidential speeches. Also, if there are parts within each document other than the president's speech, these fragments should also be deleted.

At the end of this first stage, all texts in the corpora should concern only the object of study. From there, the researcher goes through steps intended to decrease the complexity of the corpora, but without removing information important for the analysis. In this step, the focus is to change all characters to lowercase; to remove punctuation marks; and to remove the words known as stop words. All of these measures reduce the size of the document-term matrix. This matrix will serve as the basis for more complex quantitative analyses of text, such as classification methods.

The field of text analysis as data is divided into two main fields: classification method and scaling method. In this section, I approached classification methods, but specifically the three types of algorithms that allow the classification of texts: supervised, dictionary-based and unsupervised methods. The supervised model requires the existence of a set of texts that has already been codified. This set is known as training set, since its purpose is to train the classification algorithm so that it is able to classify in a more efficient and

coherent way the rest of the database that is not classified yet.

The other method, dictionary-based, as the name indicates, requires the existence of a dictionary to classify the texts. In this article, I have covered three different dictionaries: the General Inquirer Dictionary; the Regressive Imagery Dictionary; and the Linguistic Inquiry and Word Count Dictionary. Each of these dictionaries was constructed with a specific substantive preoccupation, ranging from classifying texts as either positive or negative (General Inquirer), to capturing more abstract and cognitive aspects in texts (LIWC).

Finally, the unsupervised classification method differs from the previous two because it does not require any predefined classification. This is the model that least requires a specific structure of the corpora, but it also demands a more careful analysis of the results, since there are no parameters of comparison from which the algorithm can learn.

The other way of analyzing a set of texts, which was not addressed in this chapter, is located in a continuum. Rather than sorting texts into categories, where the purpose is to assign each document into a "box," the scaling method positions each document in a spectrum. For example, we can classify texts ideologically, from the leftmost text to the rightmost one. In this method, we can also use either supervised or unsupervised procedures.

In short, all these new computational and methodological techniques have opened up a vast ocean of textual data to be explored. Once the abovementioned substantive and methodological precautions are taken, the new techniques of quantitative analysis of text enable the analysis of previously unexplored sets of texts.

3 International Attention on Brazilian Newspapers

For a long time, academics have held a debate on whether media coverage can influence public opinion or not. While some argue that newspapers can have an impact on public behavior and attitude (Iyengar and Kinder, 1987; Nelson et al., 1997; Gilens, 1999; Kellstedt, 2000), others are more incredulous (Druckman et al., 2011; Enns, 2014) and considerate that newspapers rather reflect the public's preferences and perceptions (Hopkins et al., 2017).

Whether they are a predictor or a mirror of public opinion interest, the analysis of media coverage, especially newspapers, is a fundamental tool to understand society in democracies. The importance of newspapers is shown by the work of Roberts, McCombs and others, who found that newspaper articles preceded television news coverage (Roberts and McCombs, 1994; Blood and Phillips, 1995). Therefore, they are an important source of information about the issues that are being debated or will be debated by the society.

This ability of the media, especially print media, to reflect what society is debating in a given moment can serve to test what became known as the "Almond-Lippmann consensus." This consensus emerged from the work of Almond (1950) and Lippmann (1955), where the authors stated that public opinion: (a) is unstable; (b) is unstructured (there is no coherence); and (c) does not impact the formulation of foreign policy.

Bearing this is mind, this work aims to analyze whether the Brazilian media, as a proxy of public opinion, follows the first two points of the Almond-Lippmann consensus. If the media is found to accompany international issues that on both a consistent and routinary manner over time, we will be able to affirm that the media is stable and structured with respect to international issues, and therefore as regards public opinion as well. This finding would give empirical support to the literature that states that public opinion has a greater impact on the formulation of foreign policy than predicted by the Almond-Lippmann consensus (Risse-Kappen, 1991; Holsti, 1992; Soroka, 2003a; Holsti, 2004).

Hence, in this paper I intend to analyze the international section of two Brazilian newspapers, *Folha de S. Paulo* and *O Estado de S. Paulo*. Such choice is grounded by the fact that they stayed among the top five major newspapers in Brazil throughout 2000 to 2015.[11] The other newspapers that appear in the top five were omitted from the analysis

[11]Data on newspaper's circulation can be found at: `http://www.anj.org.br/maiores-jornais-do-brasil/` – Last access: July 12, 2016.

either because they are tabloids (*Super Notícia*, *Zero Hora* and *Extra*) or because I could not have access to their articles (*O Globo*).[12] After collecting the international news, I applied a non-supervised topic modeling, LDA, to this news set. The results of the model indicate that the Brazilian media is stable, since it follows the same theme over a long period of time, and coherent, since it publishes news on topics related to international events when these events occur.

This article is structured as follows: first, I make a bibliographical survey addressing the main and most recent works on quantitative analysis of text applied to newspaper news. In the second section, I describe the process of collecting news from *O Estado de S. Paulo* and *Folha de S. Paulo*, as well as the step-by-step process to transform newspaper news from text to quantitative data. In the third section, I give a brief description of the newspaper news corpora.[13] Next, in the fourth section, I describe the topic model used to analyze this news corpora. In the fifth section, I discuss the results obtained. Finally, in the sixth section, I make the final considerations and conclude the chapter.

3.1 Media Slant in the World

The qualitative analysis of content applied to newspaper news has existed for decades, and so has the discussion of which are the best procedures to carry it out (Mintz, 1949; Stempel, 1952; Jones and Carter, 1959). Sadly, the elevated cost and the need for skilled labor to conduct research in this area has normally limited these studies to large research centers. However, the increase in computer processing capacity and the improvement in text analysis algorithms has enabled the use of the quantitative text analysis (QTA) in the exploration of a vast amount of documents without the need to hire human encoders. In this line, the work of Zhao et al. (2011) stands out for comparing traditional media to social media, specifically Twitter, by applying LDA topic modeling.

The objective of Zhao et al. was to compare traditional media with Twitter using topical modeling. This comparison is interesting because it evaluates if the topics discussed in the mainstream media are discussed in Twitter as well, but also if they are discussed as frequently in both. Another factor that increases the need for comparison between traditional media and Twitter is the format of texts itself. While texts published in traditional media –in the article, the authors consider The New York Times as traditional

[12] In the next interaction of this research, I intend to add news published by *O Globo* as well.
[13] A corpus is a collection of documents and a corpora is a collection of corpus.

media–, are long and full of details, in Twitter communication is made by micromessages, that is, messages limited to a few dozen characters, and therefore, is quite short and straightforward.

Despite this distinction in form, users of the Twitter platform access it to obtain and share news, the same way consumers of traditional media do. Therefore, Zhao et al., p.339 asked themselves "how the information contained in Twitter differs from what one can obtain from other more traditional media such as newspapers". In order to answer this question, the authors went through steps similar to those discussed in previous sections of this paper.

First, the authors collected the textual data on Twitter and the news published by The New York Times between November 11, 2009 and February 1st, 2010. Once the textual data was collected, Zhao et al. removed: stop words; the less frequent words, that is, those appearing in less than 10 documents; and the very frequent words, which appeared in more than 70% of the documents. After these changes, the authors' corpora had 1,225,851 tweets and 11,924 news stories published in the NYT. With the corpora ready, the authors ran the LDA model with 100 topics in the NYT news corpus and with 110 topics in the corpus of tweets.

To reach this amount of topics (k), Zhao et al. used some techniques to find out how many topics there were in the corpora they were analyzing. At this point, one of the main details of the topic model with LDA comes in. Given the unsupervised nature of LDA, it is a technique that requires very little from the structure of the information being analyzed; however, one of the only information the LDA model does require is the number of topics (k) that exist in the corpus or corpora.

Meanwhile, the process of discovering the number of topics of the corpus of tweets was a bit complex, since the authors were working with three hypotheses: (a) each tweet is a document – Standard LDA (nomenclature given by the authors); (b) all tweets from the same author were treated as a single document - (Author-Topic); and the Twitter platform is treated as a document (Twitter-LDA).[14] In the tests developed by Zhao et al., hypothesis three (Twitter-LDA) has presented more consistent results.

Thus, the initial LDA model used 100 topics in the news and 110 topics treating the Twitter platform as a document (Twitter-LDA). After cleaning the topics by removing

[14]Therefore, each hypothesis of the corpus of tweets was tested with 110 topics, amounting to 330 topics.

the ones that contained incoherent words (noisy topics) or common words (background topics), the authors found that there were 81 topics in the Twitter corpus and 83 topics in the NYT news corpus. With the number of topics (k) calculated and the tweets and news items allocated to their respective topics, the main finding of the article, which dialogues with this work, was that Twitter users have a low interest in international affairs, as compared to the quantity of international news published by the NYT (Zhao et al., 2011). Additionally, the authors confirmed that it is possible to use the LDA model to analyze news in traditional media.

Following the same direction, Jacobi et al. (2016) also confirmed that the use of LDA to categorize newspaper news is valid and produces consistent and coherent results. To achieve this goal, the authors designed a research to validate the qualitative results obtained by Gamson and Modigliani (1989). In this research of the late 1980s, Gamson and Modigliani analyzed the relationship between media discourse and public opinion about the use of nuclear energy.

The authors' objective was to qualitatively verify how the United States media discourse (ABC, CBS, NBC, *Time*, *Newsweek*, *U.S. News* and *World Report*) on nuclear energy had evolved between 1945 and the 1980s, since public opinion surveys showed a drop in the approval of this type of energy in the country. To that end, Gamson and Modigliani classified hundreds of predefined news and schemas, such as the *progress* schema: "Underdeveloped nations can especially benefit from peaceful uses of nuclear energy," "Nuclear power is necessary for maintaining economic growth and our way of life" and "Nuclear power opponents are afraid of change" (Gamson and Modigliani, 1989, p.11).

The authors also classified the news within the following schemas: *runaway* (fatalist); *public accountability*; *not cost effective*; and *devil's bargain*. The final result of the work of Gamson and Modigliani (1989) was an analysis of the changing occurrence of these mental schemes in U.S. media, while the most frequent one in the mid-1940s was the *progress* schema. However, the authors show that over time and with the occurrence of disasters in nuclear power plants –Three Mile Island in 1979 and Chernobyl in 1986–, *public accountability*, *runaway*, and *devil's bargain* schemas began to become more frequent than the *progress* schema in the news published by U.S. media.

These findings were instrumental in showing the connection between media and public opinion. However, the cost and time required to perform this research were considerable,

since Gamson and Modigliani had to hire and train researchers to code the news published in the 50 newspapers analyzed by them. These coders were separated in pairs, but classified the news independently. The results of the coding would only be used in the research if homogeneity between the classifications by these pairs exceeded 80%. While it is true that this measure increased the reliability of results, it also increased the costs and time of the research.

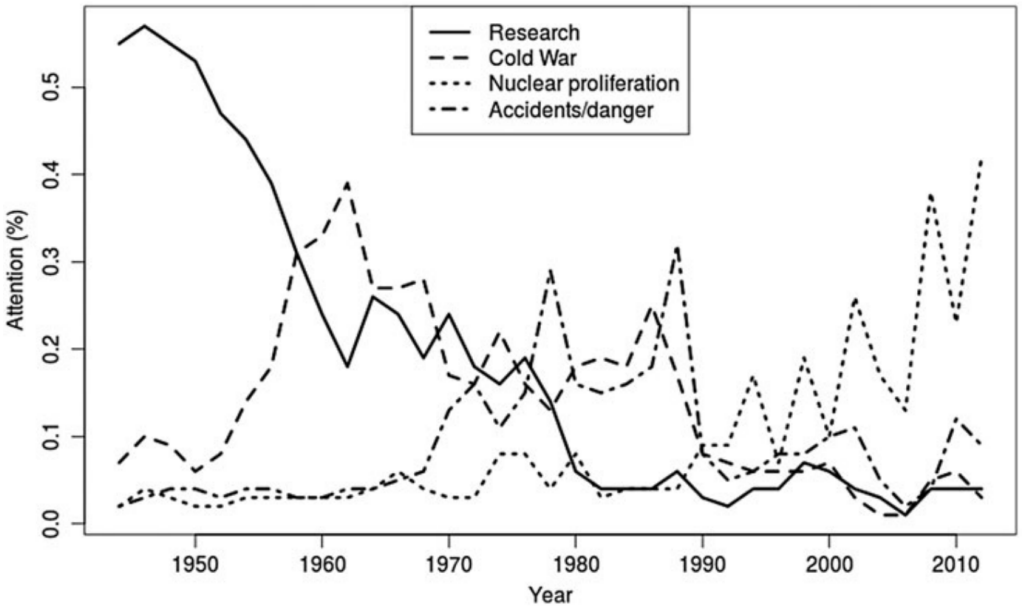
Trying to counter the time and money obstacles, Jacobi et al. (2016) depart from a framework pre-validated by Gamson and Modigliani (1989) to test whether the LDA algorithm could be used to categorize newspaper news. In order to keep the research comparable, Jacobi et al. collected all the news published in *The New York Times* (NYT) between 1945 and 2013 that dealt with the topic of nuclear energy.[15]

The final corpus had 51,528 news articles and went through a preprocessing by tokenization and lemmatization. In general lines, the first process transforms the documents into a list of words, while the lemmatizing procedure reduces the words to their lemmas. Both steps reduce the dimension and therefore the complexity of the data.[16] After these procedures, Jacobi et al. (2016) ran an LDA model with 10 topics ($k = 10$), since this was the same number of topics in the Gamson and Modigliani (1989)'s study. Figure 3 shows the results obtained by Jacobi et al. (2016):

[15]The search for these articles was done in the NYT online archives (`https://developer.nytimes.com`). The search selected news items that contained the following terms in the title or subtitle: "nuclear," "atom" or "atomic" (Jacobi et al., 2016, p.4).

[16]I will explain these two steps, tokenization and lemmatization, in more detail in the section on Topic Modeling.

Figure 3: LDA model applied to NYT's article about nuclear power



Source: (Jacobi et al., 2016, p.10)

As we can see, Jacobi et al. (2016) obtained results very similar to those found in a qualitative way by Gamson and Modigliani (1989). The topic *Research*, which is comparable to the *progress* schema, has declined over the years. This topic, however, is not the only one indicating similarities between the results of the two surveys, but also the topic *Accidents/danger*, which had the highest frequency between the late 1970s and the late 1980s. This is the final period analyzed by the study by Gamson and Modigliani, during which the authors encounter an increasing number of news in the schema of *devil's bargain*, due to the accidents in Three Mile Island and Chernobyl.

Therefore, Jacobi et al. showed that it is possible to apply the LDA model to the analysis of newspaper news, which begets economic gains, since there is no need to hire and train coders. Besides the economic advantage, the results of the LDA are replicable. On the other hand, the authors warn that the results of the LDA model should be analyzed by researchers who know the object of study substantively. This is because, as it is an unsupervised method, the LDA results may be incoherent and have no internal or external validity.

Bearing these caveats in mind, but also having the validation that it is possible to use the LDA model for the analysis of newspaper news, in the next section I will discuss the process of obtaining newspaper articles that I collected to form the international news

corpora published by *Folha de S. Paulo* and *O Estado de S. Paulo* between 2000 and 2018.

3.2 Obtaining the data

As the objective of this research is to analyze the news published in *Folha de S. Paulo* and *O Estado de S. Paulo*, the first step was to investigate whether there was a database already consolidated with such news. The main base available in the market is NexisLexis[17], but when I went deeper into its availability of data, I found that there is a great disparity in the provision of news articles for different countries.[18]

In comparison with U.S. newspapers, the NexisLexis database lacked a significant amount of Brazilian publications. In the U.S. case, the NexisLexis database contains around 270 newspapers. Not only the number of publications is noteworthy, but also the coverage of the base, which collects publications from local newspapers such as the Jupiter Courier, from the city of Jupiter – Florida[19], to world-renowned newspapers such as *The Washington Post*, *Los Angeles Times* and *The New York Times*. Also, in chronological terms, this storage of content has been done since 1980 for U.S. newspapers.[20] Thus, the NexisLexis base has a very representative sample of the U.S. print media universe, covering a significant timespan, with almost 40 years of records.

As regards Brazilian newspapers, NexisLexis does not collect news from such a significant sample of national newspapers.[21] Moreover, the period covered by this collection is much smaller as compared to U.S. data. This last point has the greatest negative impact on the quantitative text researches in Brazilian newspapers when using NexisLexis.

In quantitative terms, NexisLexis accompanies 59 Brazilian publications that issue general news in Portuguese. Currently, *Folha de S. Paulo* (Folha) and *O Estado de S. Paulo* (Estadão) are among these publications. However, back in 2015, when this research began, only *O Estado de S. Paulo* was collected by NexisLexis. This highlights the challenge of making historical studies with digitalized news from newspapers in Brazil. NexisLexis

[17] NexisLexis was founded in 1970 with the goal of providing computer-assisted legal research (CALR). More recently, in addition to this service, NexisLexis began collecting public records. In fact, in 2006 it became the largest digital database in the world in terms of legal information and public records.

[18] `http://academic.lexisnexis.eu/` – Last access: May 04, 2019.

[19] A city with around 65 thousand inhabitants.

[20] However, NexisLexis's ability to aggregate many publications in one place is beginning to draw criticism from newspapers. According to some U.S. newspapers, NexisLexis would be selling newspaper content to third parties, such as press review companies. In this way, NexisLexis would be appropriating the media companies' license agreements. https://www.thestreet.com/story/14250084 – Last access: May 04, 2019.

[21] Publications included in the NexisLexis database can be consulted at: `https://w3.nexis.com/sources/` – Last access: May 04, 2019. There, the researcher can filter results by country and type of publication.

started to collect the news from Estadão on September 2, 2009 and only started doing so with Folha news on January 18, 2018.

Therefore, and since the purpose of this research was to analyze the publications of both Brazilian newspapers, I have used two strategies, one for each publication. On the one hand, as the NexisLexis database contained the news of *O Estado de S. Paulo*, I wrote a script in R to collect all international news published by such newspaper between 2009 and the end of 2017 available on the NexisLexis website. On the other hand, since Folha's news were not on NexisLexis, I wrote another script in R to collect the news about international topics directly from the newspaper's website, covering the period between 2000 and the end of 2016.[22]

Another distinction that we must make between the two newspapers is with regard to the section that contains the international news. Table 4 lists and compare the daily sections of the two newspapers. International news are generally published in section *Mundo* [World] in Folha and *Internacional* [International] in Estadão. Although I recognize that international news can also appear in other sections of newspapers, I have collected only articles published in the international sections, where, by design, most of world news can be found.

Table 4: Newspaper sections comparison

Folha de S. Paulo	**O Estado de S. Paulo**
Mundo [World]	*Internacional* [International]
Poder [Power]	*Política* [Politics]
Cotidiano [Daily Life]	*Brasil* [Brazil]
Esporte [Sports]	*Esportes* [Sports]
Mercado [Market]	*Economia* [Economy]
Ilustrada [Culture]	*Cultura* [Culture]
Ciência [Science]	

Since news from Estadão were available on NexisLexis, it was much easier to write a script to do web scraping on the Nexis page. Because the purpose of this base is to collect and allocate the news in one place, the news therein are already standardized. Thus, the biggest challenge was to understand how the news were structured and to create a specific

[22] I've tried to write a web scraping script to collect the news directly from Estadão's webpage, so that I could not only keep the same collection strategy for the two databases, but also keep the collection pattern in the primary source. However, several features of Estadão's webpage make it difficult to create a web scraping script, such as the impossibility of searching for specific periods (for example, from January, 1st to December 31st, 2018). This possibility exists only for Estadão's archive webpage (`https://acervo.estadao.com.br/`). However, there comes the second barrier that hindered the direct web scraping of Estadão: the existence of a paywall more restricted than the one in Folha. Due to these characteristics I chose to collect the international news from Estadão through the NexisLexis portal.

script for this structure. Despite the practicality of the NexisLexis framework, it was necessary to insert several waiting times in the code so that the requests did not overload the NexisLexis servers. Because the goal was to extract a large amount of news published by Estadão, a web scraping script could end up affecting the services of a website.

The structure of NexisLexis is fairly standardized. For most of the news, it was possible to obtain information about:

- Publication date (`news_date`)
- Name of the article's author (`news_author`)
- Title of the news article (`news_title`)
- News content (`news_text`)
- Name of the section in which the news article was published (`news_section`)
- Topic of the news article (`news_topic`)
- Language of publication (`news_language`)
- Number of words in the news article (`news_length`)
- Newspaper's code in the NexisLexis system (`news_code`)
- Date the article entered the NexisLexis system (`news_load_date`)

The web scraping script to extract the news published in Folha's section *Mundo*, on the other hand, was much more complex. Since Folha's website, is not designed for data collection and analysis, as was the NexisLexis website, there is a variation over time of the location of the news on the site, but also a variation of the position of the information of interest (news title, author, content) in Folha's news pages.

As for the first variation, for example, in the 2000s, Folha's news URLs (Internet page address) started with the *http* protocol, but over time Folha gradually adopted the security protocol Hypertext Transfer Protocol Secure (*https*), which is the http protocol over an SSL/TLS protocol layer, which encrypts the connection between the server and the client. As the first version of the web scraping script I wrote only collected news from the http addresses, with the passing of the time series (from January 1st, 2000 to December 31st, 2016) the amount of news collected was declining. To understand what the error was, it

was necessary to know very well how the search engine works on the Folha webpage and what changes the mechanism had suffered over the years.

Another challenge in capturing news from the Folha webpage concerned the changing position of the information of interest in the news pages. The main issue at this point is that the script collects all information from the Folha's page where news is, and, naturally, much of this information and content is not part of the news. Figure 4, for example, is a news from the Mundo section published on October 9, 2015.[23]

[23]This article can be found at: `https://www1.folha.uol.com.br/fsp/mundo/235887-em-moscou-acao-na-siria-vira-propaganda.shtml`

Figure 4: An example of a Folha's news webpage

The news article is titled is "In Moscow, action in Syria becomes propaganda," and was written by Fernando Canzian. However, the scraping web script initially captures the entire content of the page, including the page header, with links to the other sections of the newspaper, but also all the content below and to the right of the news (print search, calendar with previous editions). Thus, it was necessary to write a script that was structured enough to filter out all these elements and leave only news and information related to it, but that at the same time was adaptable to theme and content variations

(e.g. news with images).

In the end, it was possible to extract much of the substantive news information and exclude unrelated content. The Folha database has the following information:

- Publication date (`news_date`)
- Name of the article's author (`news_author`)
- Title of the news article (`news_title`)
- News content (`news_text`)
- Name of the section in which the news article was published (`news_section`)

As the objective of this research is to analyze the corpora of news on international subjects in the two main Brazilian newspapers, the existence of similar variables in the two databases –Folha and Estadão– was essential. Therefore, this phase of obtaining the information was well completed, since all substantive variables are present in each corpus. In the next section, I will do a descriptive analysis of the two databases.

3.3 Descriptive analysis

Analyzing data in a descriptive way is an important process before any quantitative analysis, because, before building sophisticated and complex models, it is necessary to check the characteristics of the database. Some of the questions we must ask ourselves at this stage are: are the news texts well structured and "clean"? Or do they have non-news content? How does the amount of news change over time? Is there a lot of missing data on the base?

The importance of this stage is even greater in this work for two reasons. First, the collected data was obtained from two different source: the NexisLexis website, in the case of the *O Estado de S. Paulo* news, and directly from the *Folha de S. Paulo* website. Second, regardless of the source from which the data was extracted, the entire collection process was automated. Therefore, any error in the code or change in the layout of the news pages could cause errors in the database.

After several steps of sanity check, in which verification tests are done to the base content, I was able to verify that the web scraping scripts had collected the data correctly. After these steps, I created data cleansing scripts to homogenize the two databases, so all

information was left with the same pattern. For example, the format of the publication date variable (`news_date`) is year-month-day (YYYY-MM-DD) in both databases.

Once the data was verified for accuracy (sanity check) and standardized and structured (data wrangling), I proceeded to analyze the descriptive characteristics of the two databases. The first step was to verify the existence of missing data. Since the main variable of these two databases is the one containing the textual information of the news (`news_text`), my initial concern was to check if there were many cases of blank news or missing data. However, only one observation in the Estadão database had missing data in this variable, out of a total of 41,652 news items collected from NexisLexis. In the Folha database, there was no missing data in this variable. Therefore, the final database was left with 132,863 articles from Folha and 41,652 from Estadão, constituting a corpus of 174,515 news.

Regarding the temporal distribution of news, Figure 5 shows that although the search was done between January 2000 and December 2018, the database does not have information for the whole period. The news from *O Estado de S. Paulo*'s section *Internacional*, as I mentioned above, only became part of the database of NexisLexis in September 2, 2009. This phenomenon can be clearly seen in Figure 5, since the curve of the Estadão base for the average amount of news per year begins at the end of 2009. In those first months there are few news in the base of Estadão, but as early as in 2010 there was a jump and for that year the base has almost 8 thousand news from Estadão's section *Internacional*.

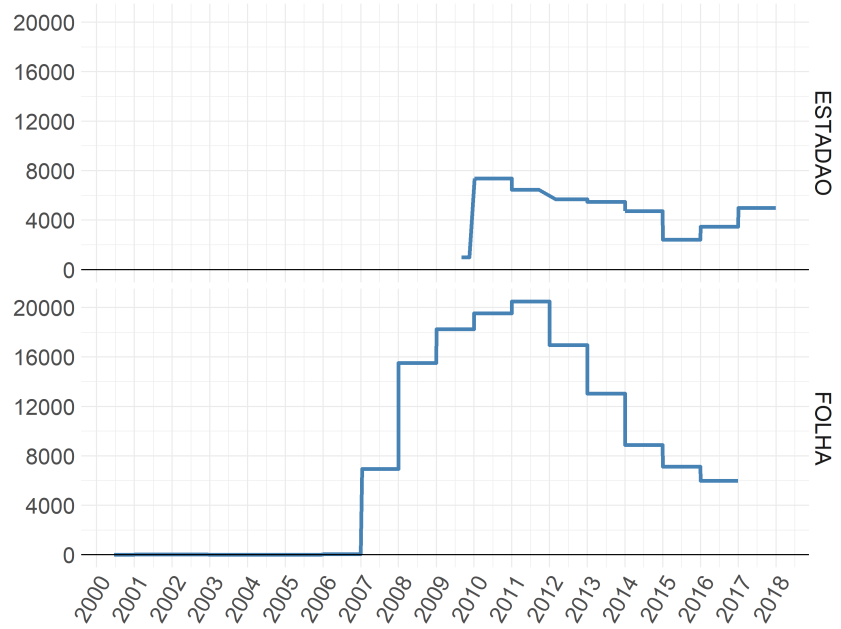
As for Folha's section *Mundo*, the database has generally more news than the base of Estadão. In addition, the time period is also longer for the Folha corpus, totaling 10 years of news –from January of 2007 to December of 2017. Another important data that can be observed in Figure 5 is that Folha's news only began to be recorded in the site's search system in January 2007. Therefore, between January 2000 and December 2006 there are no news in the corpus of Folha de S. Paulo.

Moreover, Figure 5 reveals that the international section shrank after 2012 in both Estadão and Folha, even if the drop in international news was more significant in the latter. In the peak year, Folha published an average of 20 thousand news in the international section, while in 2017 that figure went down to around six thousands news.[24] Even though

[24]I could not find an explanation for the drop in the number of international news over time in both databases. However, as this is a quantitative text analysis, if the scenario is that the news ceased to be posted on Folha's or Estadao's websites and over time and there was not a bias on the type of news being left out of the analysis, the topic modeling, which will be discussed in the next chapter, should not have

Estadão started with a lower average than Folha –eight thousand news articles in 2011–, the downward trend reversed in 2016. In 2017, *O Estado de S. Paulo* published around five thousand articles in the international section.

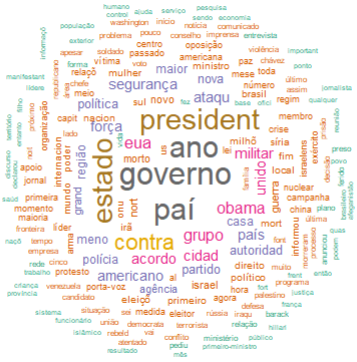
Figure 5: Average number of articles per year



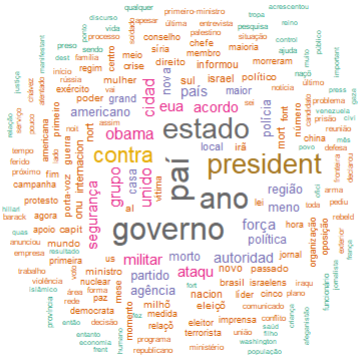
Another descriptive analysis that can be done is a word cloud plot. Figure 6 shows three word clouds; Figure 6a uses the entire dataset of news articles; and Figure 6b and Figure 6c use articles only published in Folha and Estadão, respectively. Given that there are more Folha's articles in the database, Figure 6a and Figure 6b are fairly similar. However, regardless of that, the most common words in the two newspapers are very similar. Some of the most used words in international news in the two newspapers are "government," "country" (which can be "country" or "countries"), "president" (which can be "president" or "presidents"), "state" and "year." The similarity of most used words was expected, given that the themes and period analyzed in the two newspapers are the same.

had statistical loss with this decrease, since, at the end of the time series, thousands of international news were being published per year.

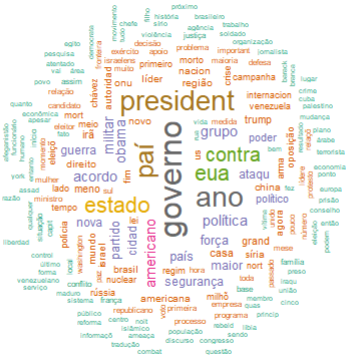
Figure 6: Word Cloud Plots



(a) Both newspapers



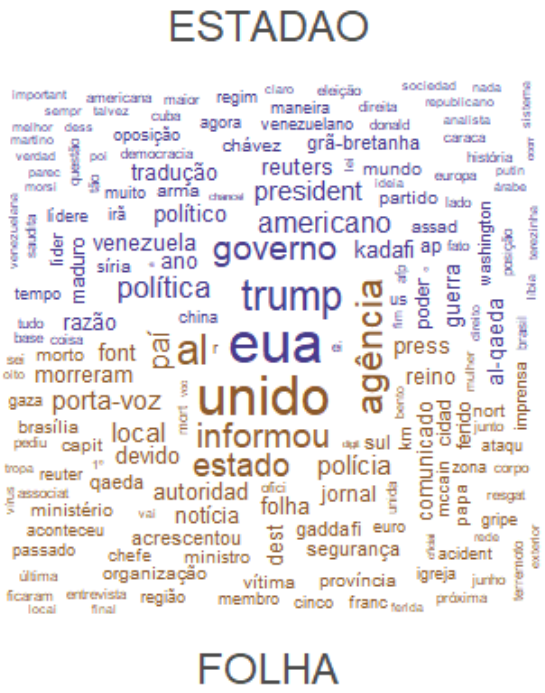
(b) Folha de S. Paulo



(c) O Estado de S. Paulo

Because it has word clouds of the most frequent words, Figure 6 allows us to see the absolute frequency in both corpora. On the other hand, it also hides the differences between the two databases, which are difficult to perceive in this world cloud configuration. That it is why I have plotted a different word cloud in Figure 7 highlighting how different Folha and Estadão are in terms of words they use.

Figure 7: Comparison word cloud plot of *Folha de S. Paulo* v. *O Estado de S. Paulo*



In Figure 7, words in blue are more used by Estadão, and words in red by Folha. The first difference between them is how they mention the United States of America: Folha tends to use "United States", and Estadão the abbreviation "USA". Another difference is how the two newspapers spell the name of former Libyan leader, Muammar Gaddafi. While Folha tends to use the version adopted by major international media outlets in English (The Washington Post, The Times, The Financial Times and The Guardian), Estadão opted to use a spell that sounds closer to Portuguese: Kadafi.[25] Finally, according to Figure 7, Estadão published more news about Venezuela. Words such as *Venezuela*, *Venezuelan*, *Chávez* and *Maduro* are disproportionately more used in Estadão's articles than in Folha's.

In sum, the descriptive analysis showed that the data are well structured and only contain the information related to newspaper news. Moreover, this exploratory data analysis (EDA) showed that the period in which there are more news about international subjects in the database is between 2007 and 2018.

Finally, both the word cloud figures with the most frequent words and the word cloud with the words that appear disproportionately more in one newspaper than in the other showed that there is not much difference between the two newspapers, at least as far as

[25] *Los Angeles Times* was one of the few U.S. newspapers to spell Kadafi.

the frequentist analysis of the bag of words is concerned, since the two publications use very similar words.[26] As the descriptive analysis showed that the two corpora are well structured, I put the two corpora together into a single corpus. In the next section I will describe how I used topic modeling to analyze this corpus with almost a decade of international news published in the two major Brazilian newspapers.

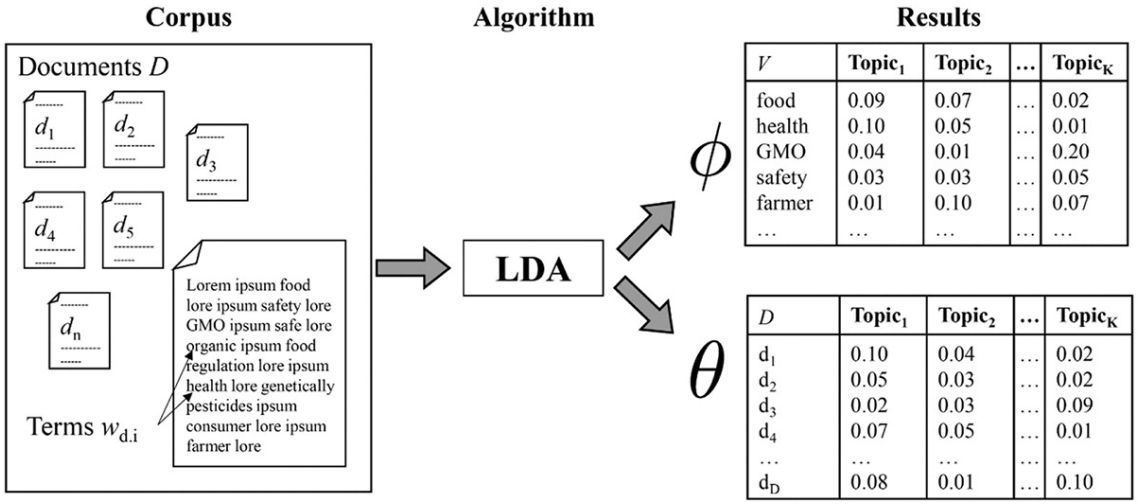
3.4 Topic Modeling

What international topics do Brazilian newspapers publish news about? How do the frequency of these themes vary over time? In order to answer these questions, in this section I will describe the topic model known as the Latent Dirichlet Allocation (LDA), which I used to analyze the international news corpora published in *Folha de S. Paulo* and *O Estado de S. Paulo*. Even though this approach had already been used by some scholars to classify news in other countries (Zhao et al., 2011; Jacobi et al., 2016), this is a very innovative initiative in the field of quantitative text analysis on international issues in Brazil.

As for the model used in this research, I chose LDA because it is a way to find topics in an unsupervised way. This is because, as the corpora have not been analyzed previously, there are no examples of pre-classified cases that could serve as an example for the supervised model to "learn" to classify. Figure 8, prepared by Maier et al. (2018), illustrates very well how the LDA algorithm works. One of the assumptions of the LDA model is that each document of the corpora is composed by several topics, and each topic is composed by a set of words.

[26] A caveat must be made at this point. As I am treating the two corpora as bags of words, that is, the order of words is not taken into account in the analysis, there may be a difference between the two newspapers when treating corpus as bigram (when the analysis is done by pairs of words: "white house" instead of "white," "house") or trigram (analysis by trios of words: "war on terror" instead of "war," "on," "terror"). These other types of treatment allow us to verify the adjectivation of nouns, as, for example, it may be that both Folha and Estadão have written extensively about former Venezuelan leader Hugo Chávez: in a bag-of-words analysis by unigram we would see "Hugo" and "Chavez"; however, in a bigram analysis, we would see that while Folha uses "dictator Chávez," Estadão uses "president Chávez."

Figure 8: Example of how LDA model works when applied to a corpus



V	Topic$_1$	Topic$_2$	…	Topic$_K$
food	0.09	0.07	…	0.02
health	0.10	0.05	…	0.01
GMO	0.04	0.01	…	0.20
safety	0.03	0.03	…	0.05
farmer	0.01	0.10	…	0.07
…	…	…	…	…

D	Topic$_1$	Topic$_2$	…	Topic$_K$
d_1	0.10	0.04	…	0.02
d_2	0.05	0.03	…	0.02
d_3	0.02	0.03	…	0.09
d_4	0.07	0.05	…	0.01
…	…	…	…	…
d_D	0.08	0.01	…	0.10

Source: (Maier et al., 2018, p.94)

Therefore, the LDA model assumes that topics can coexist within each document, so each document can contain more than one topic, and each word can belong to more than one topic. As these two assumptions exist, the result of the model also produces two pieces of information. As we can see in the right part of Figure 8, the model produces two important results: the impact each word has on each topic; and the occurrence percentage of each topic in each document.

These assumptions of the LDA model closely resemble the structure of newspaper news, since every news item can contain more than one topic. For example, one story may refer to the Crimean War, but at the same time discuss the impact of such war on Europe's trade. In addition, the word "growth" may be associated with the topic of a country's economic growth, but it may also be linked to the increasing crime in a region of the globe.

However, because it is an unsupervised model, the researcher must be cautious and verify that the results have internal and external validity. The internal validity of the result means that the topics produced contain words that have meanings correlated to an interpretable theme. For example, the words "banana," "apple" and "grape" are linked to the theme "fruit." However, if the model produces topics with words similar to "car," "paper," and "sky," it is not possible to find an interpretable label for that topic. Regarding external validity, for some researchers it is the ability of a topic model to capture events external to it (Newman et al., 2006; Evans, 2014). For example, if, when analyzing news stories that were written between September 8 and 12, 2001 in *The New York Times*, the

topic model does not create a topic for terrorist attack, it means that the model has no external validity.

Therefore, and considering the characteristics of the corpora and the lack of pre-classified examples, the unsupervised model of LDA is useful for me to analyze the quantity and content of the topics that exist in international news published by Brazilian newspapers. However, before I analyze the data, it is necessary to go through some stages of corpora preprocessing. In the next section I will describe each of these steps.

3.4.1 Preprocessing the corpora for the topic model

Before applying the LDA model, the data was preprocessed as follows. First, I created a corpus for each dataset. Then, I tokenized both corpora, by removing all numbers, punctuation marks, symbols and stop words. In relation to the stop words, I used the *quanteda* implementation for Portuguese. However, the list is not exhaustive, so I also removed words by using a customized list of stop words.[27]

Third, I applied a lemmatization algorithm to reduce variation of the same word. At this stage, works usually choose between two techniques: stemming or lemmatization. Stemming, which is the simplest technique, reduces the words to their stem. For example, both "argued" and "arguing" have "argu" as their stem. Even though the procedure by stemmer, such as the Porter Stemmer algorithm (Porter, 1980), is simpler and produces interpretable results when analyzing texts in English, for languages that have more inflections, such as Portuguese (e.g. verb "to go" in the first person of the present – *vou*, and in the first person of the future – *irei*) and German, there are little interpretable results. Therefore, it is advisable to use lemmatization algorithms instead when analyzing texts in Portuguese (Haselmayer and Jenny, 2014).

Lemmatization, like stemming, serves to reduce the amount of terms/words in a corpus. This process reduces the terms to their respective lemmas, that is, to their dictionary form. For example, the lemma of "argued" and "arguing" is "argue." Another point that

[27] Words that were also removed: "é", "ser", "nesta", "neste", "nestas", "nestes", "outro", "outros", "outra", "outras", "após", "depois", "ainda", "desde", "ter", "segundo", "desta", "dois", "afirmou", "disse", "sobre", "dia", "dias", "todo", "todos", "durante", "onde", "parte", "mil", "caso", "semana", "semanas", "três", "um", "quatro", "pode", "cerca", "ontem", "hoje", "último", "pessoas", "pessoa", "vez", "vezes", "apenas", "deve", "devem", "enquanto", "sido", "duas", "havia", "diz", "antes", "além", "segunda", "terça", "quarta", "quinta", "sexta", "feira", "cada", "vários", "várias", "domingo", "sexta-feira", "terça-feira", "segunda-feira", "alguns", "algumas", "quinta-feira", "quarta-feira", "sábado", "fazer", "porque", "sob", "têm","s", "v", "aqui", "então", "exª", "srª", "v.exª", "srs", "nº", "assim", "nesse", "sendo", "desse", "desa", "portanto", "aí", "art", "coisa", "qualquer", "quanto", "dessa", "srªs", "sras", "sr", "lá", "senhor", "todas", "tão", "nessa", "senhores", "disso", "alguma", "pois", "desses", "tendo", "sobretudo", "quais".

makes lemmatization different from stemming is that the first one takes into account the context in which the word is used: the lemma of the word "saw" may be "see" or "saw" depending on whether the word is being used as a verb or as a noun. As a consequence, for languages that have more deflected words, as is the case for Portuguese, it is advisable to use lemmatization algorithms (Haselmayer and Jenny, 2014).

After reducing the corpora's dimensionality with lemmatization, I have created a document-feature matrix (*dfm*) using both tokenized datasets, which is a format of data that the *quanteda* package in R uses to analyze text. Then, I was able to convert both corpora to a document-term matrix (*dtm*), which is the data format that the LDA command accepted. The structure of the *dtm* is quite simple: each document is in a line and each term (word) is in a column, and the cells receive information on how many times that term occurs in each document.

Once the *dtm* is built, the LDA model only needs one more information, which is the amount of topics (k) in the corpora. This is the only information that must be input by the researcher. In the next section, I will discuss the methodological and substantive procedures I used to find the value of k.

3.4.2 Number of topics: $k = 80$

Because it is an unsupervised model, *a priori* the LDA model presupposes very little about the structure of the analyzed data. In fact, the only information the researcher must provide to the model is the number of topics (k) that exist in the corpora. One possible approach that I could have taken was to use the number of topics that LexisNexis uses to classify most of *O Estado de S. Paulo*'s news articles. The issue with this path is that there are over 350 topics, most of them with just one or a couple of articles.

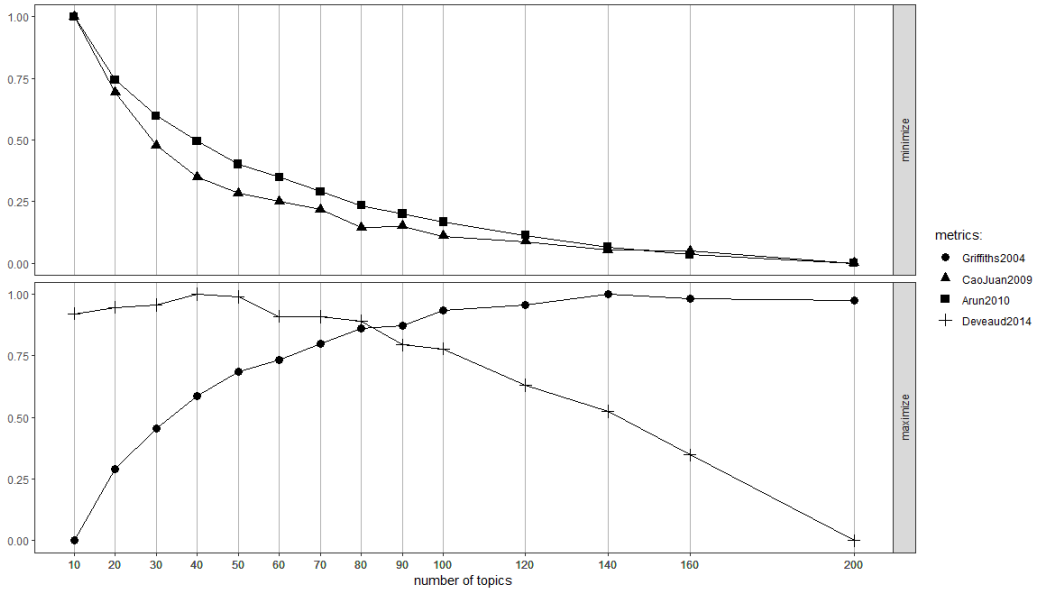
Therefore, $k = 350$ seems to be a very high value. Therefore, I chose to use two quantitative strategies to choose the value of k. The first procedure involved analyzing four different metrics and how they behaved by changing the number of k to: 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 120, 140, 160, 200. Figure 9 was built with the *R* packet called *ldatuning*, by using the `FindTopicsNumber` command.[28]

The optimal value of k should be the one that maximizes *Griffiths2004* (Griffiths and Steyvers, 2004) and *Deveaud2014* (Deveaud et al., 2014), but minimizes the values of

[28]The computer took 72 hours with parallel processing to run this command.

Arun2010 (Arun et al., 2010) and *CaoJuan2009* (Cao et al., 2009). As we can see in Figure 9, the optimum value of k is 80. This is because, after this value, the *Deveaud2014* metric begins to fall, and the *Arun2010* and *CaoJuan2009* values minimize only marginally.

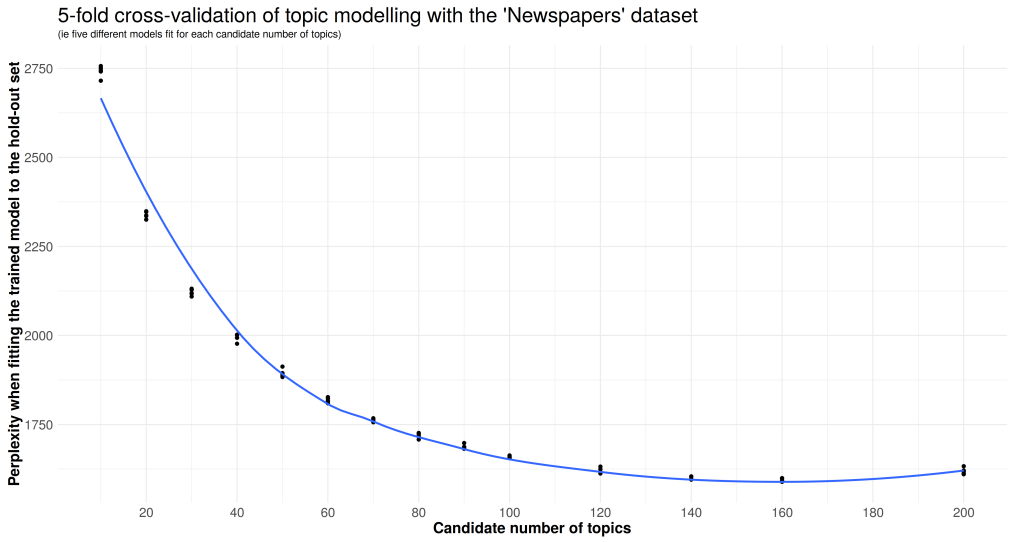
Figure 9: Number of topics



Another strategy I used to determine the value of k was to evaluate the perplexity value, which should be as small as possible. Figure 10 shows the perplexity value for different values of k: 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 120, 140, 160, 200.[29] For each value of k, I run five LDA models with different fits. The perplexity value decreases well until $k = 60$, but after $k = 80$ the reduction is quite marginal.

[29]To build Figure 10, I followed the commands created by Nidhi: `http://www.rpubs.com/MNidhi/NumberoftopicsLDA`

Figure 10: Number of topics: Cross-validation



Both strategies showed that the optimal value of k is 80.[30] Therefore, with the *dtm* constructed and the value of k stipulated at $k = 80$, I applied the LDA model in the corpora of international news published by Folha and Estadão in the last decade. In the next section, I will detail the results of the model, which were consistent and had internal and external validity.

3.5 Results: What and when – International news in Brazilian newspapers

The results of the LDA model in the international news corpora of the two main Brazilian newspapers show that the range of subjects is quite diverse, but there is a strong concentration around issues involving conflicts and wars. As we will see in this section, the results obtained through the topic model are consistent and have internal and external validity. The findings are consistent not only in terms of the top keys that influence each topic, but also regarding the way each topic behaves in a time series perspective.

Table 11, in Appendix, shows the list of topics and the corresponding keys that influence them. For each of the 80 topics there is a list, in descending order, of the 30 words that most impact the topic formation. For example, the topic on Libya (Topic 2) has as the 10 most significant words for the topic's creation the following terms, in descending order: "libyan," "gaddafi," "rebel," "dictator," "city," "Tripoli," "regime," "force," "country,"

[30] Even though, I chose to run the topic modeling with $k = 80$, I also ran the model with k equal to 40 and 60. In Appendix A, I listed the results for those ks and labeled each of the categories generated.

"Kadafi."[31] In this case, the term "libyan" has the most impact on topic creation. After this term, the word "gaddafi" and then "rebel" are the most influential in creating the topic on Libya.

Coherence and internal validity of the results can be observed through the analysis of each of the topics and the connection between the main terms that create each topic. In the case of the topic "Libya," these terms, listed above, are directly interconnected with each other. It is important to remember that the analyzed corpora has articles between 2007 and 2018, a period in which Libya went through two civil wars, having been controlled until 2011 by political leader Muammar Gaddafi. Gaddafi, or Kadafi, was considered a dictator and died in 2011, 8 months after the beginning of a civil war that began with civil protests for the overthrow of his regime. The following conflict between forces loyal to the regime and rebels was characterized by the conquering of a city at a time, and was mainly determined by who controlled the capital of the country, Tripoli. Thus, the ten terms are quite consistent with each other and are directly linked to the issue of Libya.

Coherence and internal validity also occur in the vast majority of other topics. In cases where there was no coherence between the main terms that form the topic, I labeled the topic as "Unknown." There are five topics where there is no clear coherence. For example, in the topic "Unknown 1," the ten most influential words to create the topic were "local," "brasilia," "fire," "city," "train," "time," "hour," "bus," "fire" and "fireman." Or topic "Unknown 4," where the top ten terms were "political," "problem," "question," "clear," "time," "fact," "great," "important," "difficult" and "moment." In neither of these two topics is there a clear coherence between the words listed that could be assigned to a specific theme or topic. Thus, we can use the nomenclature of *noisy topic* and *background topic* (Zhao et al., 2011), and state that the LDA model produced 75 background topics, which contain words related to a theme, and 5 noisy topics, that is, topics that contain incoherent words.

Regarding the proportion of news articles per topic, Figure 11 shows that the distribution of topics is more evenly spread when compared to the political speeches' topic distribution. Also, the five most common topics published by Folha and Estadão are about: Israel–Palestine; Syria–Lebanon; Natural Disasters; U.S. primary elections and Iran. Meanwhile, the five topics with the least amount of articles are: Year–Month, Un-

[31] Given that the corpora underwent a process of lemmatization, terms like "Líbia" [Libya] and "líbio" [Libyan] became just "líbio" [Libyan].

known 2, Unknown 4, Diplomacy and International Negotiation.

Figure 11: Proportion of news articles per topic ($k = 80$)

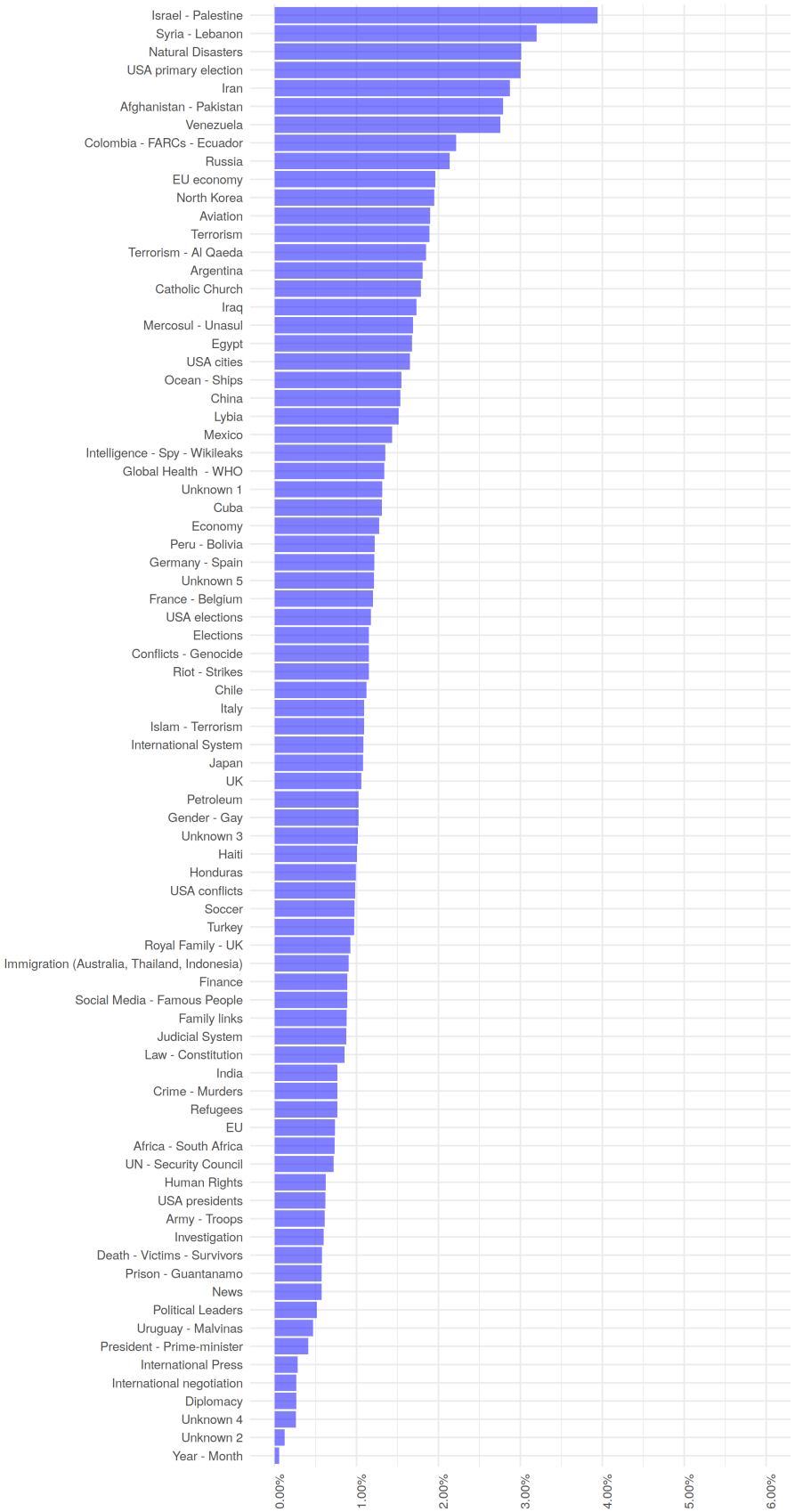


Figure 11 is also an interesting source for validation of the model, in addition to the analysis of which words most impact the formation of each topic. This is because, among the five topics that had less news stories assigned to them, three do not have substantive content, namely: Year–Month, Unknown 2 and Unknown 4. Also, the most frequent noisy topic is Unknown 1, and ranks 27^{th} among the most frequent. Therefore, the noisy topics are not very significant in quantitative terms given the model used.

Another relevant fact that can be abstracted from Figure 11 is the indication that Brazilian newspapers when dealing with international issues have a greater bias to conflict than to peace. This is because the two most frequent topics (Israel–Palestine and Syria–Lebanon) are clearly related to armed conflict, while Diplomacy and International negotiation –the two less frequent topics if we disregard non-substantive topics– are more linked to peace journalism (Galtung, 2003; Lee and Maslog, 2005; Keeble et al., 2010).

As we have seen so far, the results of the LDA model have internal validity, since words that impact the formation of each topic are interconnected with each other and to an identifiable topic. The next step is to verify whether the results also have external validity. The strategy available to identify external validity is to observe the temporal pattern in which the topics occur and compare this pattern with real events that took place during the period analyzed (Newman et al., 2006; Evans, 2014).

Next, I will use graphs to illustrate the temporal behavior of the proportion of published news. In addition to showing the external validity of the model, it also helps highlight the substantive findings of the result. In order to do so, I will focus on four topics, which cover three very different themes: the first two graphs are on the topics "U.S. elections" and "U.S. primary elections;" the third chart contains data from the topic "Global Health – WHO;" and, finally, I will describe the results for the topic "Russia." All graphs indicate that the LDA model with 80 topics produced substantial results with external validity.

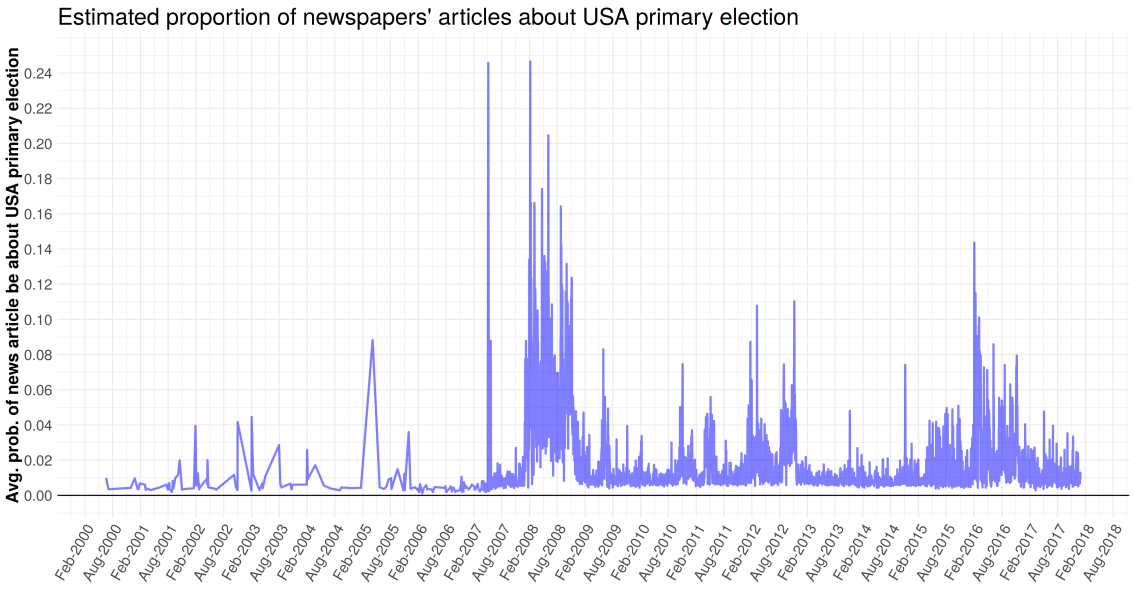
First, as expected –as it is the largest economy on the planet–, the United States is the main country represented within the 80 topics. There are five topics that are directly related to the U.S., namely: "U.S. primary elections," "U.S. cities," "U.S. elections," "U.S. conflicts" and "U.S. presidents." As we can see, the topics range from hard power ("U.S. conflicts") to soft power ("U.S. cities") themes. Another interesting point is the attention that the Brazilian media gives to the political environment in the United States, especially to the Executive Branch. Particular attention is paid to the president's election processes,

from the primary elections stage –where voters from the two main parties, the Democratic Party and the Republican Party, choose their candidates for the general election– to the general election, held every four years.

Figures 12 and 13 illustrate these two steps. On the one hand, Figure 12 indicates that the topic about U.S. primary elections being captured is the dispute inside the political parties. These disputes occur a few months before the month of November, the month in which the presidential elections take place. On the other hand, Figure 13 depicts how the presidential general election topic developed over time in the newspapers.

Since presidential elections occur every four years, in the period analyzed they occurred in 2000, 2004, 2008, 2012 and 2016. However, since there is little news published before 2007 in the database, we can only assess the impact that the 2008, 2012 and 2016 elections had in the model. In 2008, when Barack Obama was first elected, Folha and Estadão extensively covered the U.S. elections. During that year, Figure 12 reveals that the proportion of articles published in the international section about that topic got over 25%, which means an average of one out of four articles in the international section being about the primary elections in the U.S. As results accompany events that occurred in real life and are directly related to the topic theme, Figure 12 attests the external validity of the model.

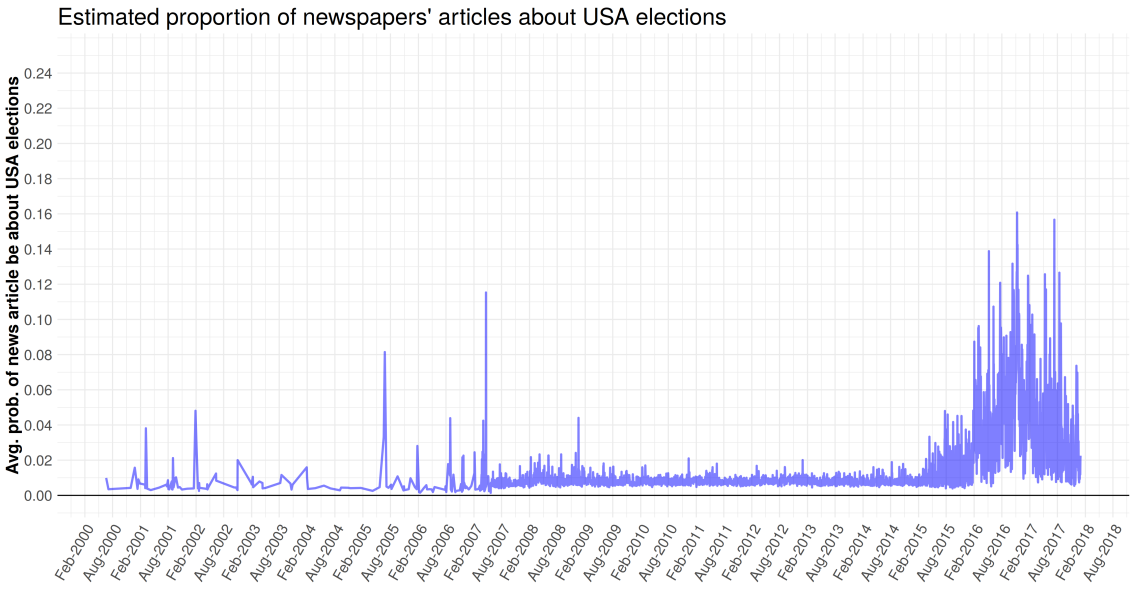
Figure 12: Proportion of news articles about Primaries in the US elections



As mentioned earlier, in addition to external validity, the results also bring substantial

gains. Figure 13, combined with the results in Figure 12, shows that normally, after the primary elections, Brazilian newspapers do not cover the presidential elections in the United States to the point of creating a new topic. However, sharp political polarization in 2016's presidential elections between candidates Donald Trump (Republican Party) and Hillary Clinton (Democratic Party) (Gentzkow, 2016; Boxell et al., 2017; Allcott and Gentzkow, 2017) caused the topic "U.S. elections" to capture practically only this last election. Following a fierce dispute, Trump was elected by electoral college votes (Trump: 304; Hillary: 227), but would have lost if the popular vote was considered directly (Trump: 46.1%; Hillary: 48.2%).

Figure 13: Proportion of news articles about U.S. Elections



The behavior of the topic on Global Health (Global Health – WHO) also allows validating the results both internally and externally. The internal validity can be observed in the analysis of the ten most influential terms for the formation of the Global Health topic, listed here in order of decreasing influence: "doctor," "illness," "case," "hospital," "flu," "virus," "country," "death," "treatment" and "swine."

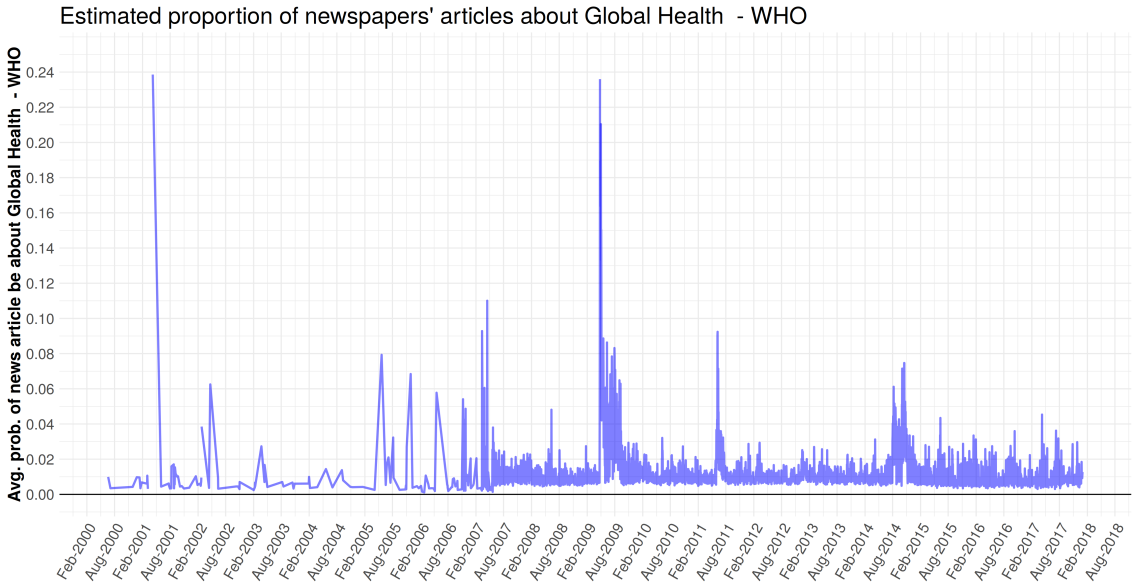
All these terms are related with diseases ("illness," "flu," "virus"), but also with the notion of territory ("country") and healing capacity ("doctor," "hospital," "treatment"). In order to have a better understanding of these relationships, as well as to verify the external validity of the model, Figure 14 illustrates the temporal evolution of this topic

in proportional terms in the corpora. When we disregard the years prior to 2007, given that little news has been collected for that period, we can see that there are four peaks of occurrence of the topic Global Health – WHO: in 2007, 2009, 2011 and 2014.

In all these years, there was a case of potential pandemic or acute crisis in terms of global health. In 2007, the case of lead in paint used in Chinese toys caused millions of toys to be recalled.[32] In April 2009, the pandemic of Influenza A spread around the world, so much that the WHO's pandemic alert level reached its maximum level (6) on June 11, 2009. Because of its swine origin, this pandemic was also known as the swine flu (term that appears among those that most influence the topic) or Mexican influenza, since its outbreak was in Mexico.

In 2011, the outbreak of E. coli O104:H4 was in Germany, but originated from Spanish cucumbers. A total of 16 countries recorded thousands of cases of infection by the bacteria. Finally, between 2014 and 2016, the outbreak of Ebola devastated the western part of the African continent, infecting around 30,000 people and killing more than 11,000. The precarious infrastructure of treatment and diagnosis of the countries in Western Africa made it difficult to contain this epidemic. It ended up spreading to at least seven countries (Liberia, Mali, Nigeria, Senegal, Sierra Leone, Italy, Spain, the United Kingdom, and the United States) after the first case was discovered in Guinea.

Figure 14: Proportion of news articles about Global Health – WHO

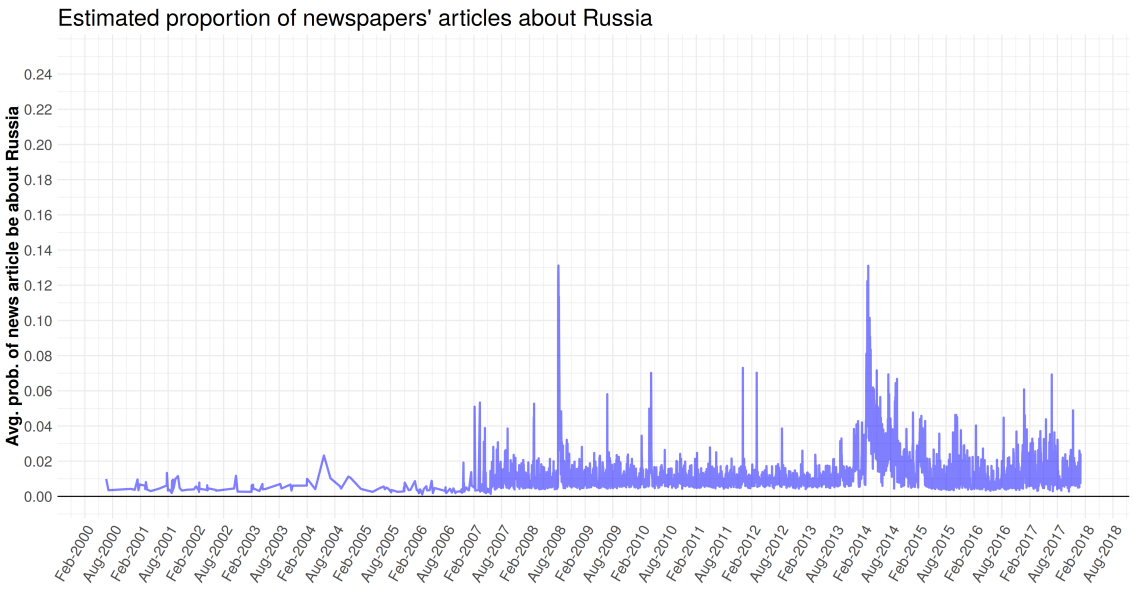


[32] https://www.reuters.com/article/us-mattel-fisherprice/fisher-price-recalling-1-5-million-toys-idUSWEN003320070802 – Last access: May 04, 2019.

Another case that also reinforces the external validity of the model is the topic "Russia," in which the ten most influential terms are: "russian," "russia," putin," "ukraine," "president," "ukrainian," "vladimir," "soviet," "separatist" and "georgia." There are two international events that occurred during the period of the corpora and which are related to Russia: the Russo-Georgian War and the Crimean Crisis, two armed conflicts involving Russia and former Soviet countries: Georgia (South Ossetia and Abkhazia), and Ukraine, in the case of Crimean Crisis. For much of that time, Vladimir Putin was the Russian president.

The Russo-Georgian War took place from August 7 to 16, 2008. In Figure 15 it is possible to clearly see the increase in published news items on the topic Russia in August 2008. The second event, Crimean Crisis, occurred between February 23 and March 28, 2014, and its effects on the amount of published news can also be observed in Figure 15.

Figure 15: Proportion of news articles about Russia



Therefore, the results of the LDA model in the international news corpora published in *Folha de S. Paulo* and *O Estado de S. Paulo* show that the model produced results with coherence, internal validity and external validity. In addition to bringing a substantial gain to the area of media analysis, by showing which themes the Brazilian newspapers publish more about and when these themes are addressed. In the next section, I will conclude the chapter and cast the future agenda based on the findings of this research.

3.6 Conclusion

The objective of this section was to analyze which international issues the Brazilian newspapers publish news about, both in static terms, that is, at a given moment, and to assess how flows of international themes occur over time. To do so, I analyzed the international news published by the two major Brazilian newspapers: *Folha de S. Paulo* and *O Estado de S. Paulo*. The period of analysis comprised years from 2000 to 2018, but much of the news collected was concentrated between 2007 and 2017.

With this set of news from two different corpora, I have structured a database to which I then applied a non-supervised topic model called LDA – Latent Dirichlet Allocation. In order to analyze the news using the LDA, it was necessary to "clean" the news database by removing the following terms: stop words, numbers, adverbs and verbs. These steps were aimed at decreasing the dimensionality of the document-term matrix (*dtm*), which is the data format accepted by the LDA command. With the *dtm* ready, I applied the LDA model with 80 topics ($k = 80$), meaning that I assumed that there are 80 different topics in the international news corpora. This value (k) was validated by cross-validation tests that minimized perplexity metrics, among other tests.

The results of the LDA model are coherent and have internal and external validity. In addition to this validation of the model, the results also mean substantial gains for the area of international relations, since it made possible to identify the prevalence of news related to great powers, European Union, Russia, China, but mainly about the United States. These findings alone were expected, given that these are the most important countries in military or economic terms. However, the temporal analyses enabled by the model indicate that the prevalence of these topics in the Brazilian media is not a static standard. Each topic accompanies the occurrence of international events related to each topic.

For example, the topic on "U.S. primary elections" has peaks occurring every four years, 2008, 2012 and 2016, from February to August. These periods correspond exactly to the periods when primary elections occur in the United States. Second, the analysis of the topic "Global Health – WHO" allows the clear visualization of when the last outbreaks of infectious diseases that became epidemics occurred. Third, the occurrence of the topic "Russia" shows the last military interventions of Russia in countries of the former Soviet Bloc, Ukraine and Georgia.

Whereas in this chapter I treated the corpora of newspaper news as a monolithic bloc, since the purpose was to analyze how Brazilian media treats international issues, my future intention is to run a topic model for each newspaper. In a second interaction of this work, then, I will treat each corpus, Folha and Estadão, as distinct entities. By doing that, we will be able to verify which international topics Folha and Estadão address the most. In addition to this analysis, we will also be able to verify whether the same topic –for example, "Elections in the U.S."– is treated more positively or negatively by Folha than Estadão.

Another improvement that I would like to incorporate to the next interaction of this research would be adding the news published in the *Mundo* [World] section of *O Globo*, since this newspaper ranks as either the second or the third biggest Brazilian newspaper, depending on the metrics used, and is published in Rio de Janeiro, which would enrich the research with a point of view outside São Paulo.

Finally, it is important to emphasize that this work creates a very interesting line of research for the area of International Relations in Brazil, for it shows that quantitative text analysis, and more specifically topic modeling, can be used to analyze international news. With the results found in this research it is possible to create an automated observatory of the Brazilian press on international issues. This observatory could monitor the publication of news in real time and allocate them into the 80 topics listed in this research.

Moreover, it would be possible to create automatic alerts that would warn when the proportion of published news passed a threshold that would be predefined for each topic. For example, if the amount of published news exceeds 5% of the average probability of news articles about "Global Health – WHO," it is very likely that a pandemic or infectious disease is occurring somewhere in the world. In short, even if there is a long way to better understand how the debate on international issues takes place in the media, we now know that we can use statistical techniques to shed light on this issue.

4 Political Speeches

Are Brazilian politicians interested in international issues? And, if so, what international issues do they speak about? Do these speeches follow events external to the dynamics of the political sphere? Or are politicians rather insular, and so debates on international issues follow the rules of procedure of the political system, remaining unaffected by external events? Bearing these questions in mind, in this paper I will analyze the speeches made in the Committees on Foreign Relations of both the Chamber of Deputies and the Senate in Brazil.

In order to achieve this goal, I have put together an unprecedented database, created specifically for this article. This basis contains all the speeches made in the two Foreign Relations Committees known as CRE (Senate) and CREDN (Chamber of Deputies), between January 1st, 2000 and December 31, 2017. During this period, 62,410 speeches were delivered in the two Committees, by both members and guests.

Besides being a sea of data not yet explored, this database allows us to advance in the analysis of how the Legislative Branch formulates foreign policy. As we will see in this article, debates on international issues in the Legislative occur more frequently than in the Executive, and they are mostly open to the public. In contrast, debates on international issues by the Executive often happen behind closed doors, and only a few pronouncements are made publicly. This scenario is especially true in Brazil, given the centrality of Itamaraty in the formulation of Brazilian foreign policy and its isolation from the political system and public debate.

After creating this database, I have applied an unsupervised topic modeling for 60 topics ($k = 60$). The findings of this research unfold into methodological and substantive contributions. In relation to the methodological findings, this work shows that it is possible to use quantitative text analysis tools to examine the set of speeches in the Brazilian Legislative. The results obtained are reliable, can be reproduced by other researchers, and have internal and external validity.

Finally, the substantive findings of this research show that members of CRE and CREDN are neither silent nor disinterested in international affairs. Among the 60 topics found and labeled, there is coherence and awareness with international topics that affect or have affected Brazil since 2000. Another substantive data obtained from the results of

the topic model is the existence of a temporal dynamics in the occurrence of topics in the Committees. This indicates that politicians are not isolated from international events, but follow and are influenced by them.

The paper is structured as follows. First, I discuss the importance of looking at the Legislative when we look at the formulation of foreign policy, and I also present the most recent works that went along this line. In the second section, I analyze the institutional design of the two houses of the National Congress of Brazil, specifically the two Committees on Foreign Relations. In the third section, I describe the step-by-step process of collecting the speeches on the websites of the Senate and the Chamber of Deputies, and I present a descriptive summary of the data collected. In the fourth section, I present and discuss the topic model used to analyze the speeches. In the following section, I present the results, and finally, in the sixth section, I make the final considerations.

4.1 Bringing the Legislative back in

Although some articles analyze the Brazilian case, from party manifesto analyses (Tarouco, 2011; Tarouco and Madeira, 2013) to analyses of politicians speeches (da Silva, 2016; Moreira, 2016), the literature is mostly focused on the political discourse in the United States (Petrocik, 1996; Kahn and Kenney, 1999; Sides, 2006; Gadarian, 2010) or, when analyzing Brazil, focuses rather on political discourse by the Executive Branch (Brandi Aleixo, 1988; Fonseca and Monteiro, 2005; da Silva, 2012).

As the objective of this research is to analyze the performance of the Brazilian Legislative Branch on international issues, we first need to differentiate it from the Executive. Kurz (1990, p.70) notes a difference between Legislative and Executive when dealing with foreign affairs, that is, the difference between open and close bureaucracy. Even though Kurz used the US system to build this comparison, we can employ the same analogy to the Brazilian case.

On one hand, while the Executive's foreign affairs bureaucracy is very opaque and shrouded in secrecy, especially if we consider the way Itamaraty handle them (Pimenta de Faria, 2008; Aurélio Pimenta de Faria, 2012). On the other hand, the Legislative is more open, given that most hearings of the Committee of Foreign Relations are open and that anyone can have access to the meetings' transcript.

Therefore, studying the Brazilian Legislative allows us to analyze a vast and rich num-

ber of documents and speeches. This possibility is further intensified given that in recent years the Brazilian National Congress has begun to digitalize and publish all documents online, including the transcription of speeches given in the plenary sessions of both Houses, or even in their numerous committees. Thus, the study of the Legislative allows us to map and analyze a greater amount of material, as compared to studies on the foreign policy of the Executive.

Regarding the literature on Brazilian Legislative and political parties, Tarouco's methodology is different from what we propose in this work. In comparative terms, both studies made by Tarouco carried out content analysis by reading excerpts from government programs, which were then scored by a researcher. A shortcoming of this method is that different encoders may assign different scores, thus diminishing scientific reliability. I propose using the Quantitative Text Analysis (QTA) method to look for similarities in the documents (press news and political speeches) and the type of words used to allocate them into topics. Given that the researcher determines the software routine, this process increases reliability.[33]

In a recent thesis, Moreira (2016) carries out a quantitative analysis of the speeches in the National Congress of Brazil. His main goal was to understand whether parliamentary speech is determined by the government-opposition spectrum. However, Moreira did not separate domestic issues from international ones in his analysis of Congress members' speeches, leaving our query about the dynamics of foreign issues inside the Legislative unanswered.

In his thesis published in 2016, da Silva's evaluates whether Congress members are "silent" in the debate on foreign policy in Brazil. The author finds that Congress members do participate in foreign policy. However, da Silva analysis does not include speeches, only Congress members' votes. In his conclusion, da Silva recognizes the importance of analyzing parliamentary speech to better understand the relationship between foreign policy and Congress behavior.

It is our understanding, therefore, that there is a gap in the literature on political speeches on international issues in Brazil. Mainly, there are no studies that systematize and analyze the speeches and oral statements made within the Standing Committees on Foreign Relations of the National Congress, CRE and CREDN. In order to better understand how

[33]The flip side of the coin for QTA is that research validity issues may occur.

these two standing committees work, in the next section, I will analyze the Brazilian Legislative Branch by focusing on the institutional design of these two Committees on Foreign Relations.

4.2 Brazilian Congress' Committees on Foreign Relations

The Brazilian institutional design was strongly inspired by the United States model (Riker, 1964; Duchacek, 1970). According to these authors, the modern federative State, based on republican principles, was invented by the United States of America. Consequently, according to this US-centered view, all subsequent experiences of republican federal States were influenced by the political-institutional structure of the United States (Arretche, 2001).

Not only has Brazil adapted the US federalist structure; it has also divided the powers into the Executive, Legislative and Judiciary branches. The principle of division of powers was conceived by Montesquieu (1748) and incorporated into the American Constitution by the founding fathers of the United States (Hamilton et al., 2005). When analyzing the Brazilian Legislative Branch, we can see other aspects thereof that were inspired by the institutional design of the United States. Like the United States, Brazil has a bicameral legislature, divided into *Câmara dos Deputados* (Chamber of Deputies) and *Senado* (Senate). While the first one is intended to represent the Brazilian population with its 513 seats[34], the 81 senators represent the different federal units.[35]

However, the bicameral structure is not the only similarity with the U.S. legislative system. Just as in the U.S., the Brazilian Congress, besides having plenary sessions, a body in which all deputies of both houses can vote and express themselves, the work is split among dozens of Committees.

The Committees are thematic and may be permanent, temporary or mixed (with members of both Legislative Houses). The Committees act in two fronts: legislation and supervision. The legislative functions are based on the discussion and elaboration of laws within their respective thematic areas. The second group of functions regard the

[34] As the Chamber of Deputies has the role of representing the population, and it is not evenly distributed throughout the national territory, each state votes in a number of deputies ranging from 8 to 70, depending on its population. The most populous state, São Paulo, chooses 70 deputies, and the least populous state, Roraima, has eight seats in the Chamber of Deputies. The discussion of overrepresentation or underrepresentation exists (Nicolau, 1997; Soares and Lourenço, 2006), but will not be addressed in this work.

[35] Each of the 27 units of the federation elects three senators.

ability of the Committees to supervise the actions of the Executive branch linked to the thematic area of each Committee. Finally, the Committees also have the role of hearing and debating different issues with the civil society.

Therefore, given these responsibilities and the thematic division, a more detailed debate on bills is carried out at Committee level. Committees can review bills in a conclusive way, or else forward them to the Plenary. In the first case, the Committee, in assessing the matter in a conclusive manner, may approve or reject bills, and this decision is made effective.

In the second case, the Committee chooses a rapporteur, who shall give an opinion on the impacts of the bill. The opinion is then approved or rejected by the members of the Committee. If the opinion is approved, the bill is sent to the Plenary for a vote. The importance of the rapporteur's opinion, besides guiding the debate within the Committee, lies in the fact that it will also guide the plenary when it comes to deciding on the matter. Article 24, subsection II, of the Internal Rules of Procedure of the Chamber of Deputies sets forth what types of projects are conclusive and which ones should also be voted by the Plenary. As a general rule, projects that modify constitutional rights should be submitted to evaluation by the plenary.

Given the importance of the Committees, it is relevant to understand how their members are chosen. This process is very similar in both Legislative Houses. In the Chamber of Deputies, Article 10, subsection VI of the Internal Rules of Procedure, establishes that party leaders must indicate to the house's management board which representatives from their political groups will compose the Committees. The board then proceeds to allocate the deputies by observing proportionality and representativeness of each political party. In a similar procedure, the President of the Senate, upon recommendation of the party leaders, appoints the members of the Committees.

However, this appointment is not exempt from the pressures of the ruling party. According to Pereira and Mueller, "*[the] Executive, through party leaders or a governing coalition, manipulates the appointments of certain committees in order to have in them a strategic number of members that is faithful to their interests*" (Pereira and Mueller, 2000, p. 49). In this sense, the Committees' ability to verify and limit the performance of Executive power, a procedure known as checks and balances, may be limited.

Given their responsibilities, Committees' meetings are an important source for under-

standing not only how the Brazilian Legislative participates in the checks and balances mechanism, but also how each politician, representing a political party, discusses the matters handled by the Committees. Thus, this research seeks to advance the knowledge we have about the Legislative by analyzing the behavior of Congress members in these bodies. However, since the main purpose of this research is to draw an analysis of such behavior over time, my query is limited to standing committees.

Table 5 lists all standing committees in the two legislative houses. The first noteworthy datum is the difference between the number of standing committees in the Chamber of Deputies compared to those in the Senate. In the Chamber, there are 25 standing committees, that is, excluding temporary committees and mixed committees (composed by both deputies and senators), while in the Federal Senate there are 14 standing committees. The comparison of absolute values is, nonetheless, unfair, since the Chamber of Deputies has 513 seats and the Federal Senate has only 81. Thus, in a relative comparison, there are more committees in the Senate than in the Chamber of Deputies. There is a committee in the Senate for every 5.8 senators, while in the Chamber of Deputies there is one for every 20.5 deputies.

Although the Brazilian Congress has several committees, the most important standing committees that deal with foreign affairs are: *Comissão de Relações Exteriores e de Defesa Nacional* (CREDN) in the Chamber of Deputies, and *Comissão de Relações Exteriores e Defesa Nacional* (CRE) in the Senate, both roughly translating into "Committee on Foreign Relations and National Defense." My analysis will focus on these two committees, due to time span – they are standing committees – and the thematic they deal with. In the following sections, I will describe how each committee on foreign affairs works in each legislative body.

4.2.1 Senate's Committee on Foreign Relations

The Committee is composed by 19 senators and their respective alternates. Nonetheless, the Committee's size may vary, for senators can invite special guests, such as ambassadors, professors, bureaucrats and Ministers of State. In general, these guests are summoned to explain world issues and their consequences to Brazil. Since the Senate's Committee on Foreign Relations is also responsible for confirming the ambassadors appointed by the president, the period analyzed herein includes several Senate confirmation hearings.

Table 5: Brazilian Congress' Standing Committees

Standing Committees	
Chamber of Deputies	**Federal Senate**
Comissão de Agricultura, Pecuária, Abastecimento e Desenvolvimento Rural (CAPADR)	Comissão de Assuntos Econômicos (CAE)
Comissão de Ciência e Tecnologia, Comunicação e Informática (CCTCI)	Comissão de Assuntos Sociais (CAS)
Comissão de Constituição e Justiça e de Cidadania (CCJC)	Comissão de Constituição, Justiça e Cidadania (CCJ)
Comissão de Cultura (CCULT)	Comissão de Ciência, Tecnologia, Inovação, Comunicação e Informática (CCT)
Comissão de Defesa do Consumidor (CDC)	Comissão de Direitos Humanos e Legislação Participativa (CDH)
Comissão de Defesa dos Direitos da Mulher (CMULHER)	Comissão Diretora do Senado Federal (CDIR)
Comissão de Defesa dos Direitos da Pessoa Idosa (CIDOSO)	Comissão de Desenvolvimento Regional e Turismo (CDR)
Comissão de Defesa dos Direitos das Pessoas com Deficiência (CPD)	Comissão de Educação, Cultura e Esporte (CE)
Comissão de Desenvolvimento Urbano (CDU)	Comissão de Serviços de Infraestrutura (CI)
Comissão de Desenvolvimento Econômico, Indústria, Comércio e Serviços (CDEICS)	Comissão de Meio Ambiente (CMA)
Comissão de Direitos Humanos e Minorias (CDHM)	Comissão de Agricultura e Reforma Agrária (CRA)
Comissão de Educação (CE)	Comissão de Relações Exteriores e Defesa Nacional (CRE)
Comissão do Esporte (CESPO)	Comissão Senado do Futuro (CSF)
Comissão de Finanças e Tributação (CFT)	Comissão de Transparência, Governança, Fiscalização e Controle e Defesa do Consumidor (CTFC)
Comissão de Fiscalização Financeira e Controle (CFFC)	
Comissão de Integração Nacional, Desenvolvimento Regional e da Amazônia (CINDRA)	
Comissão de Legislação Participativa (CLP)	
Comissão de Meio Ambiente e Desenvolvimento Sustentável (CMADS)	
Comissão de Minas e Energia (CME)	
Comissão de Relações Exteriores e de Defesa Nacional (CREDN)	
Comissão de Segurança Pública e Combate ao Crime Organizado (CSPCCO)	
Comissão de Seguridade Social e Família (CSSF)	
Comissão de Trabalho, de Administração e Serviço Público (CTASP)	
Comissão de Turismo (CTUR)	
Comissão de Viação e Transportes (CVT)	

Beside this function, the Senate's Committee on Foreign Relations has the following legal responsibilities and duties:

> **Article 103.** The Committee on Foreign Relations and National Defense is responsible for issuing opinions on:
>
> I - propositions referring to international acts and relations (Const., Article 49, I) and to the Ministry of Foreign Affairs;
>
> II – foreign trade;
>
> III – nomination of heads of permanent diplomatic missions to foreign governments and international organizations to which Brazil is a party (Const., Article 52, IV);
>
> IV – (Revoked);
>
> V – Armed Forces of land, sea and air, military requisitions, passage of foreign forces and their stay in the national territory, issues regarding borders and limits of the national territory, air and sea space, declaration of war and celebration of peace (Const. Article 49, II);
>
> VI – matters relating to the United Nations and international organizations of all kinds;
>
> VII – authorization for the President or Vice-President of the Republic to leave the national territory (Const., Article 49, III);
>
> VIII – other matters related thereto.
>
> 1st Paragraph. The Committee shall be a part, through one of its members, of the commissions sent by the Senate abroad, in matters concerning Brazil's foreign policy.
>
> 2nd Paragraph. The Committee shall hold public hearings at the beginning of each legislative session, with the Ministers of Foreign Affairs and Defense, to provide information within the scope of their competences.
>
> (Internal Rules of Procedure of the Federal Senate)

In the evaluation of Schmitt (2011), CRE has a bigger influence over Brazilian foreign policy than the Committee on Foreign Relations and National Defense of the Chamber of Deputies (CREDN). This strength comes from the third clause of Article 103 of the Internal Rules of Procedure of the Senate: [CRE is responsible for] "appointing the head of permanent diplomatic missions to foreign governments and international organizations." Assessment sessions are held so that senators can evaluate the experience and knowledge of the diplomat who is applying for a position as head of a permanent diplomatic mission. After the session, senators must vote for the approval or rejection of the candidate.

In May 2015, even after CRE approved the diplomat Guilherme Patriota, brother of former Foreign Minister Antônio Patriota, as head of the Brazilian mission to the OAS, the Senate floor repealed his nomination. In a tight vote, 38-37, the plenary rejected for the first time in the history of Brazilian politics a diplomat appointed by the Executive.

4.2.2 Chamber of Deputies' Committee on Foreign Relations

The Committee on Foreign Relations and National Defense (CREDN) of the Chamber of Deputies has 36 members (as of the beginning of the 2019 legislature). However, the number of members is not unchangeable from legislature to legislature. That is because, according to the Internal Rules of Procedure of the Chamber of Deputies:

> **Article 25.** The number of permanent members of the Standing Committees shall be determined by an act of the Bureau, after hearing the Leader Collegiate, at the beginning of each legislature.
>
> 1st Paragraph. Such determination shall take into account the composition of the Chamber in light of the number of Committees, to assure the observance, as far as possible, of the principle of party proportionality and other criteria and norms for the representation of seats.
>
> (Internal Rules of Procedure of the Chamber of Deputies)

As for the functions of the CREDN, they are focused rather on the legislative rules of procedure, as compared with the functions of CRE (Senate). The main functions of CREDN are reviewing bills and international treaties, and serving as a check-and-balance mechanism for the Executive's actions in the field of international relations. Unlike the Senate's Committee on Foreign Relations, CREDN does not interview diplomats appointed to the position of ambassador. Although it has fewer powers than its counterpart in the Senate, CREDN must also authorize the exit of the President or Vice-President from the national territory. Finally, the CREDN has the role of promoting and listening to the public debate on foreign policy issues. To this end, dozens of public hearings are convened each year, and academics, members of civil society and ministers of State are invited to discuss international topics.

In his master's thesis, da Silva (2012) empirically analyzes the CREDN between the 52nd and 53rd legislatures. The author develops his analysis in three dimensions: individual, institutional and party-wise. On the individual level, da Silva concludes that

representatives are not absent in the debate on international affairs. However, this participation does not take place in the government versus opposition axis. This finding is interesting, since many studies show that the action of legislators in the Brazilian Congress is highly explained by the opposition-government logic (Cheibub et al., 2009).

In order to analyze parliamentary behavior, the author examined the votes within the CREDN and found that there are few cases of opinion rejection. Therefore, the opposition-government dynamic is not in place, at least not strongly, within this Committee. However, the author himself makes a caveat about this finding: since the number of opinions that are rejected is very small, there could be a government-opposition dynamic in place before opinions were evaluated by CREDN. In this scenario, which was not tested by the author, the political parties of the ruling base would make a preliminary inquiry into how the members of the Committee would vote a particular opinion. If these governing parties suspect that there may be strong opposition to it, the report could be taken off the agenda. Thus there would be few cases of rejection and the opposition-government dynamic would be present.

Regarding the institutional dimension, da Silva (2012) also did not find a government-opposition relationship in the selection of chairmen and rapporteurs in CREDN. This result indicates that the pressure of the Executive is not transmitted to the choice of the two most relevant positions within the Committee on Foreign Relations, since the Chair of the Committee is responsible for:

> **Article 41.** The Chair of the Committee shall, in addition to what is assigned to them in these Rules of Procedure or in the Rules of Procedure of the Committees:
>
> II – convene and preside over all meetings of the Committee, ensuring the necessary order and solemnity; ...
>
> VI – appoint Rapporteurs and Alternate Rapporteurs and assign to them the matter subject to review (...)
>
> (Internal Rules of Procedure of the Chamber of Deputies)

Therefore, the Chair of the Committee has the power to set the agenda, and is responsible for choosing the rapporteurs. As a consequence, were the Executive to interfere in CREDN, most of the chairs would presumably be from the ruling parties. Nevertheless, during 5 out of 8 years of the 52nd and 53rd legislatures, CREDN was chaired by an opposition representative (da Silva, 2012).

Still at the institutional level, the author also did not find interference of the Executive in the choice of rapporteurs. The importance of the position of rapporteur is that the opinion produced by them rules the whole discussion at the time of voting. Therefore, at the institutional level, representatives seem to have certain autonomy in relation to the Executive.

Finally, at the party level, da Silva (2012) examines which representatives the Congress leaders have appointed to compose the CREDN, based on the deputies who were appointed to compose the Committee on the Constitution and Justice and Citizenship (CCJC). The choice of CCJC as a point of reference is justified by the consensus in the literature that it is one of the most important Committees of the Chamber of Deputies (Müller, 2005), since it can prevent bills and constitutional amendments from moving further. The results show that if we take the members of the CCJC as a parameter, members of CREDN are not very different. In terms of seniority, parliamentary expertise and party discipline, CREDN members are very similar to CCJC members (da Silva, 2012, pp. 72-73).

The findings described above indicate that the performance of deputies in CREDN is not determined by a strong presence of the Executive. In view of this, the behavior of Congress members within this committee tends to be freer than the behavior of legislators in the Chamber's floor (Pereira and Mueller, 2002).

4.3 Political speeches: collecting data

As discussed in the previous section, there are some studies that analyze the behavior of Congress members in the Foreign Affairs Committees of the Chamber of Deputies (da Silva, 2012) and the Senate (Schmitt, 2011). They seem to find similar results: Congress members are neither silent nor indifferent to international issues.

However, these analyses have focused on the meetings' minutes, votes and/or the profile of the politician (seniority or which political party the member of the Committee belongs to). Several of the tests carried out do not indicate any government-opposition relationship or parliamentary indiscipline, since negative cases are rare. As we have seen in the previous sections, such cases may be rare due to ex ante government-opposition control of Committee votes; also, the political party does not randomly select the politician to be a member of the Committee.

Consequently, analyzing votes and the profile of members of the Committee may pro-

duce inaccurate results. However, what Congress members talk about in these Committees is freer from the influence of the Executive and party leaders. The politician can vote in favor of the subject pressured by the Executive or the leader of their party. However, they may complain or address a different topic in their speech. In this sense, I am interested in analyzing the political speeches made in the Foreign Relations Committees in the two Legislative Chambers, an effort that has not been made so far.

In relation to the politicians' speeches database, one of the first challenges is data availability. Between 1964 and 1985, Brazil underwent an authoritarian regime that restricted the participation of the population in the formulation of politics. Furthermore, speeches previous to the 1990s are not fully digitalized. As a result of these two caveats, I decided to analyze all the speeches between 2000 and 2017. More specifically, I analyze the politicians' speeches given in both Committees on Foreign Affairs of the National Congress of Brazil during the period.

Speeches delivered within the Committees can be accessed as shorthand notes. Although most of the shorthand notes are available online, the extraction of each speech, and the further identification of the author, their position and/or political party, were tricky enough. Below, I describe this process, on a step-by-step basis, from the collection of shorthand notes to the separation of speeches. First, I describe how the speeches in CRE (Senate) were collected and prepared for analysis, and then I detail the procedures used to extract the speeches in CREDN (Chamber of Deputies).

With regard to CRE, I downloaded all minutes available in the Senate Foreign Affairs Committee's website[36] for meetings held between 2000 and 2017. Each meeting has its corresponding minutes file, and the file extensions vary between `*.doc` and `*.rtf`. The files amounted to 550 minutes. With the files stored, I wrote a R script to read each file and paste its content in a data frame. After that, I split the meetings' minutes per speech, thus creating a database with 44,005 lines. In this process, I also separated all data per speaker: name, meeting date, political party (when available) and profession.

Unlike the CRE, speeches in CREDN are in `HTML` format. Between January 1st, 2000 and December 31st, 2017 there were 1,147 meetings of this Committee.[37] This figure includes all types of meetings that CREDN either promoted or was a part of. This includes

[36] `http://legis.senado.leg.br/comissoes/comissao?2&codcol=54`

[37] `http://www2.camara.leg.br/atividade-legislativa/comissoes/comissoes-permanentes/credn/reunioes/pesquisa_reunioes_comissao`

different meeting formats, such as panels, symposia, general meetings, extraordinary public hearings and joint public hearings with other committees, among other types.

Since CREDN's speeches are in `HTML` format, instead of downloading the files of each speech as it was done for CRE, I wrote a script in *R* that read all the pages bearing the speeches. After collecting all these pages, the script separates each speech and identifies them by their respective authors. Finally, I wrote a script in *R* to check if the author of the speech was a political representative. After this verification, the script unites the politician with their respective political party. For cases where speakers were not politicians, I researched and inserted the profession data into the database. CREDN's final database had 18,405 lines, where each line was an uninterrupted speech made within CREDN. In the following section, I make a descriptive summary of the CRE and CREDN databases.

4.3.1 Data summary

The database of the two Committees on Foreign Relations of the National Congress of Brazil, CRE and CREDN, has 46 variables and 62,410 observations. The table with the statistical summary of all variables can be consulted in Table 12, Appendix 12. As we will see in the graphs below, it is important to point out that there is missing data in some specific periods throughout the historical series from 2000 to 2017. In CRE, the Senate Committee, there is no data for years between 2009 and 2011. In the Chamber of Deputies' Committee, CREDN, no data were found between the years 2000 and 2002.[38]

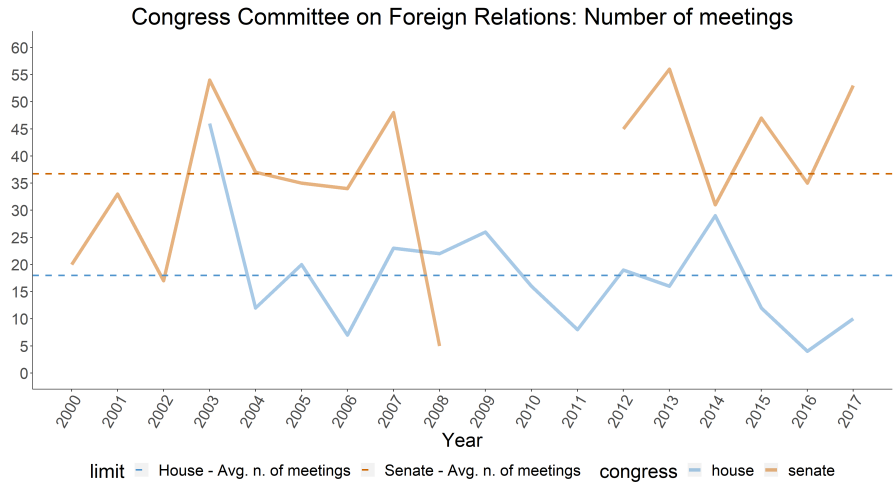
Regarding the frequency of Committees' meetings, we found that CRE meets on average approximately 37 times a year, while its counterpart in the Chamber of Deputies gathers 18 times a year. This figure includes all types of meetings: symposia, lectures, public hearings, general meetings, joint meetings (those with participation of other committees), etc. Therefore, on average, Committees' meetings take place more than once a month.

Figure 16 shows the annual and period average number of times CRE and CREDN met, dividing the total number of meetings in Figure 816a and removing joint meetings from the calculation in Figure 816b. Logically, the average meeting occurrence drops when

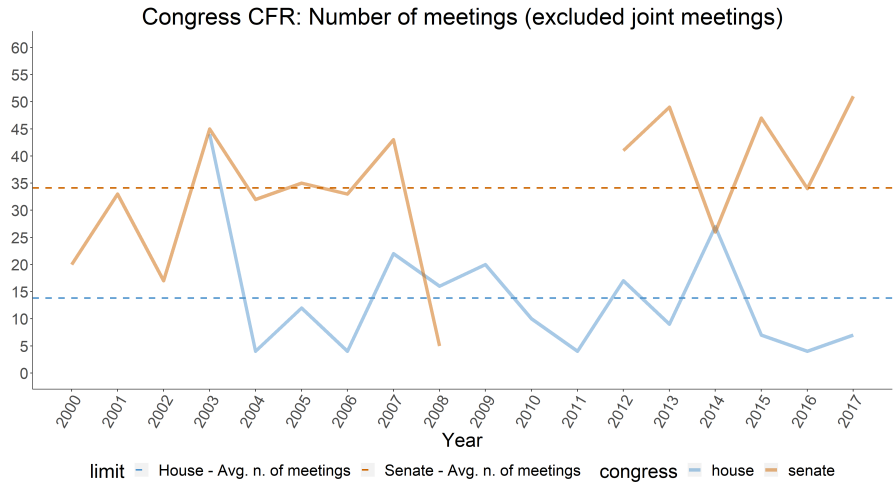
[38]The collection of speeches in CRE and CREDN was made for the entire period of analysis, from 2000 to 2017. However, there is no record of the meetings in these two periods mentioned, from 2000 to 2002 on the Chamber website and from 2009 to 2011 in the website of the Senate.

we remove joint meetings from the calculation. This decrease is proportionally higher in CREDN, which now meets on average 14 times a year, while CRE meets on average 34 times *per annum*, when we discount the joint meetings.

Figure 16: Congress CFR: Number of meetings per year



(a) Considering joint meetings

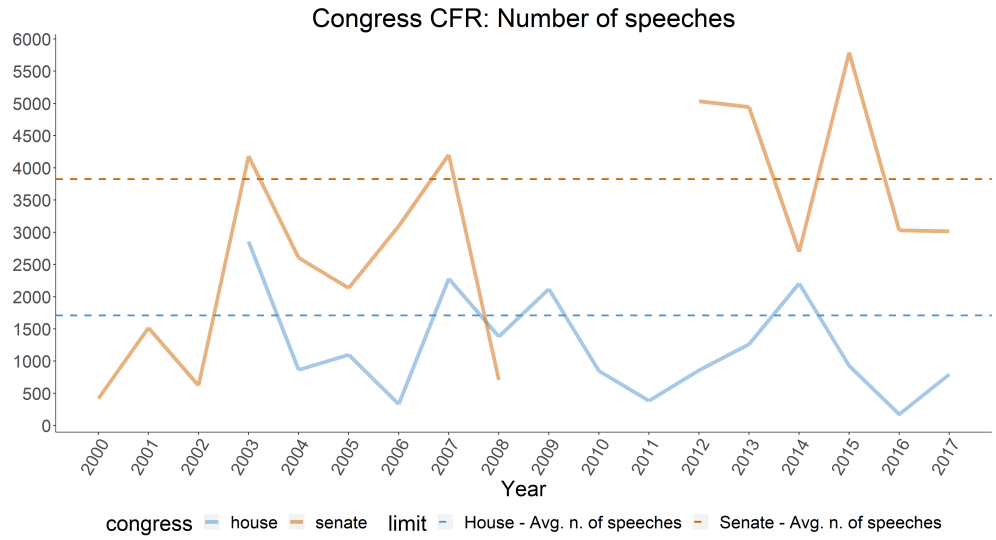


(b) Excluding joint meetings

Even if the number of speeches varied considerably over the period analyzed for both committees, the change in the number of speeches made had a different pattern when comparing CRE and CREDN. In CRE, Figure 17 highlights that there was an increase in the number of speeches over the period, jumping from 500 speeches in 2000 to 3,000 speeches in 2017, with a peak of 5,500 speeches in 2015, while in CREDN the number of speeches remained close to the average throughout the period, with no increase in the number of times CREDN members and invited speakers expressed themselves. In 2005, there were more than 2,500 speeches in CREDN; by 2017, however, this number drops

to roughly 500. Finally, we can observe that, throughout most of the time, there were more speeches delivered in CRE than in CREDN. This fact was expectable, given that, annually, there are more CRE meetings than CREDN's, as shown in Figure 16.

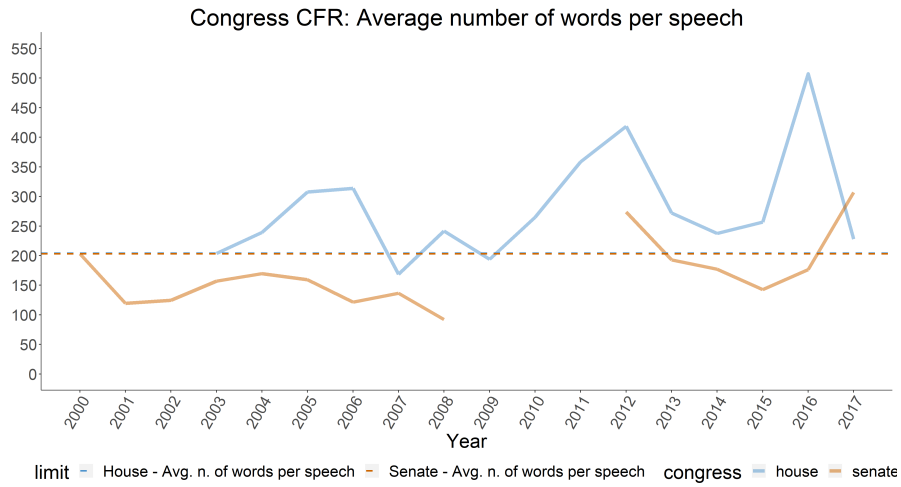
Figure 17: Number of speeches per year (Congress CFR)



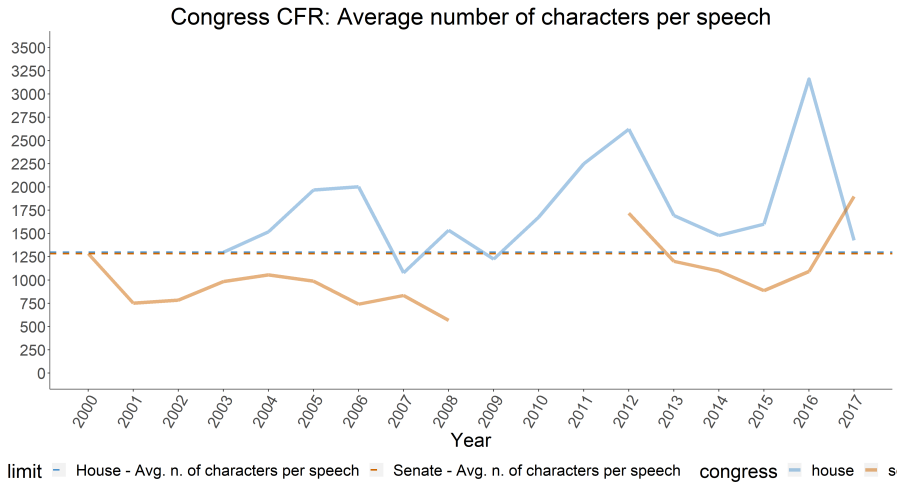
Regarding the duration of speeches, Figure 18 reveals that the change was not only in terms of absolute number of speeches, but also in relative terms. Figure 18a shows that the average number of words per speech has increased over time. In CRE, the average speech length increased by 50% during the period analyzed. In 2000, the average speech length was 200 words, and in 2017, it was just over 300 words. In terms of average number of characters per speech, the same can be said, as Figure 18b shows.

The fact that the number of characters follows the same pattern as the number of words used throughout the period indicates that the average size of words remained the same. Another noteworthy fact is that the average length of speeches is very similar in the two Committees, CRE and CREDN. Throughout the period analyzed, the average speech duration in both CRE and CREDN was a little over 200 words, with an average of almost 1,300 characters per speech.

Figure 18: Congress CFR: Average speech length per year



(a) Average number of words per year



(b) Average number of characters per year

In relation to who does most of the talking in each Committee, Table 6 and Table 7 list the 20 people that talked the most in the Senate's Committee on Foreign Relations (CRE) and in the Chamber of Deputies' Committee on Foreign Relations (CREDN), respectively, during the period analyzed.

The startling result that Table 6 points out is that the vast majority of speakers listed are not politicians. As a matter of fact, there is only one politician, Jairo Jorge da Silva (PT). The most common profession in this list is Ambassador. Even though it is surprising that the guests give longer speeches in the Senate's Committee on Foreign Relations, we find the results to be consistent with the role of the Committee.

Given that the Senate's Committee on Foreign Relations has to ratify the candidates appointed by the President of the Republic for ambassador positions, each Ambassador

undergoing evaluation has to give a speech, which tends to be long, to explain his or her opinion on several issues, usually related to the country that he or she is being assigned to. The second most common professional category when it comes to long speeches includes bureaucrats and Ministers of State.

This occurs because senators can summon State officials to clarify decisions being made or to explain the consequences of a trade agreement to Brazil, for example. Therefore, Table 6 indicates that the process of cleaning and separating speeches was done correctly. That means the results have external validity, that is, ambassadors and bureaucrats are expected to give the longest speeches given the CRE's institutional design.

Table 6: Top 20 Longest Speeches on Average – CRE (Senate)

Rank	Name	Profession	Avg. Number of Words per Speech
1	José Gregori	Minister	4291
2	Affonso Alencastro Massot	Ambassador	3692
3	Samuel De Abreu Pessôa	Economist	3688
4	Ronaldo Sardenberg	Ambassador	3047
5	Jairo Jorge Da Silva	Politician (PT)	2844
6	Vitória Alice Cleaver	Ambassador	2843
7	Adalberto Tokarski	Bureaucrat	2742
8	Luiz Augusto Saint-Brisson De Araújo Castro	Ambassador	2663
9	Nelson José Hubner Moreira	Minister	2651
10	Flávio Roberto Bonzanini	Ambassador	2503
11	Ben Boer	Professor	2503
12	Creomar Lima Carvalho De Souza	Professor	2459
13	Carlos Roberto Pio Da Costa Filho	Professor	2443.5
14	Francisco Roberto De Albuquerque	General	2409.3
15	Dante Coelho Lima	Ambassador	2388
16	Kywal De Oliveira	Ambassador	2385
17	Fausto Martha Godoy	Ambassador	2341
18	Fernando Antonio Lyrio Silva	Bureaucrat	2336.5
19	José Ricardo Roriz Coelho	Entrepreneur	2272.7
20	Pedro Luiz De Oliveira Jatobá	Entrepreneur	2259

Table 7 indicates that the procedure for cleaning and separating speeches in CREDN was also done correctly. As we mentioned earlier, it is not up to CREDN of the Chamber of Deputies to evaluate the diplomats appointed to head diplomatic missions. Thus, we should not expect a large number of diplomats to be on the list of people who made the longest speeches on average in that instance.

As shown in Table 7, there are only three diplomats, and the most common profession in this top 20 list is bureaucrat. This is because CREDN, like CRE, invites bureaucrats and experts to debate the issues on the agenda. In this sense, the number of teachers in

Table 7 deserves attention: there are five of them, the same number as bureaucrats.

Lastly, an interesting fact, also observable in CRE, is that there is only one person directly attached to a political party in the list of top 20 members with the longest speeches on average. Edialeda Salgado do Nascimento, who has an average of 3,935 words per speech, was the national secretary of PDT's Black Movement. She was also the first black woman to assume a Secretary of State in Brazil, as secretary of Social Promotion in the first Brizola Government (Rio de Janeiro, 1982).

Table 7: Top 20 Longest Speeches on Average – CREDN (Chamber of Deputies)

Rank	Name	Profession	Avg. Number of Words per Speech
1	Edialeda Salgado Do Nascimento	Politician (PDT)	3935
2	Fernando José De Camargo	Consultant (*LCA Consultores*)	3795
3	Benedito Adalberto Brunca	Bureaucrat	3585
4	Elias Khalil Jabbour	Bureaucrat	3533
5	Paulo Gilberto Fagundes Vizentini	Professor	3333
6	Pedro Ivo Batista	Environmentalist	3172
7	Marcelo Falak	Journalist	3077
8	José Carlos De Souza Braga	Professor	3064
9	William Respondovesk	Bureaucrat	3061
10	Antonio De Aguiar Patriota	Ambassador	3039
11	Jacqueline Ramos Silva Carrijo	Bureaucrat (Labor Inspector)	2844
12	Luciana Acioly Da Silva	Professor	2785
13	Javier Jordan	Professor	2747
14	Antonio José Ferreira Simões	Diplomat	2737
15	Helder Mutéia	FAO (UN)	2642
16	Fernando Paulo De Mello Barreto	Ambassador	2617
17	José Euclides Da Silva Gonçalves	Air Chief Marshal	2606
18	Creomar Lima	Professor	2526
19	Júlio César Imenes De Medeiros	Bureaucrat (FINEP)	2513.5
20	Vitor Afonso Coutinho	Helibras	2379

In summary, the descriptive analysis of the databases of the two Foreign Relations

Committees of the National Congress showed that the collection and cleaning of speeches were done correctly, since the variation in the number of meetings and speeches in the Committees indicates a correlation with reality, as discussed above. This external validation is fundamental, since the two databases are unpublished.

In addition, since the entire process of collecting, cleaning and separating speeches was done automatically, there was the possibility of errors at each step of these procedures. Therefore, after confirming that the databases are well structured, I can move on to the analysis of the actual statements.

In the next section, I will describe how the topic modeling was carried out, the step-by-step cleaning, and the analysis of the results.

4.4 Topic Modeling

It would be a Herculean task to analyze the speeches of Congress members in the Foreign Relations Committees of the National Congress (CRE and CREDN) manually. Since, as mentioned earlier, between 2000 and 2017 there were 62,410 speeches in the two Committees, reading and labeling each speech with a topic would be a work of years. Thus, I have resorted to the quantitative text analysis procedure in order to automate a significant part of the analysis process.

More specifically, in this section I will describe the use of the topic model technique to analyze the themes appearing in Congress members' speeches in CRE and CREDN. For that purpose, I will use an algorithm known as Latent Dirichlet Allocation (LDA), which I will explain later.

This section is divided as follows: I first describe the cleaning steps of the set of speeches in the two Committees, hereinafter referred to as *corpora* (plural of *corpus*); then I describe how the LDA algorithm works; next, I discuss and explain the step-by-step process that must be covered to find the number of topics in the corpora; and, finally, I present the results of the analysis of topics, showing that there is both internal and external validation of the topics found with LDA.

4.4.1 Preprocessing

Before applying the LDA model, the data was preprocessed as follows. First, I created a corpus for each dataset. Then, I tokenized both corpora, removing all numbers, punctua-

tion marks, symbols and stop words. In relation to the stop words, I used the *quanteda* (a *R* package) implementation list of Portuguese stop words. Since this list is not exhaustive, I also removed words using a customized list of stop words.[39]

Third, I applied a stemming algorithm to reduce variation of the same word. After that, I created a document-feature matrix (*dfm*) for both tokenized datasets. I was then able to convert both *dfms* into a document-term matrix (*dtm*), which is the data format that LDA command accepted. The following section focus on the method used to analyze the topics debated in both Committees, CRE and CREDN.

4.4.2 Methodological procedures

In order to understand whether the two Committees on Foreign Relations of the National Congress discuss the same issues over time, I decided to use topic models. This approach had already been used for political speeches by several scholars (Quinn et al., 2006; Zirn and Stuckenschmidt, 2014; Greene and Cross, 2015; Gautrais et al., 2017). Nonetheless, given the novelty of the corpora analyzed herein, this method will shred light in a sea of data that has been hidden so far.

More specifically, the technique I used to obtain the topics was Latent Dirichlet Allocation (LDA) (Blei et al., 2003). A major advantage of LDA is being able to automatically assign documents into topics. The only information that the researcher has to provide is the collection of documents (N) and the number of topics (k).

This feature is particularly important due to the dimension of the corpus analyzed herein: when combined, there are 62,410 speeches in the Committees.[40] Moreover, this corpus has not been previously classified into different topics. As a result of such, I was not able to use a set of topics from a preceding research as a training set for my analysis. Therefore, given that these datasets have not been classified yet, I used unsupervised

[39] Words that were also removed: "é", "ser", "nesta", "neste", "nestas", "nestes", "outro", "outros", "outra", "outras", "após", "depois", "ainda", "desde", "ter", "segundo", "desta", "dois", "afirmou", "disse", "sobre", "dia", "dias", "todo", "todos", "durante", "onde", "parte", "mil", "caso", "semana", "semanas", "três", "um", "quatro", "pode", "cerca", "ontem", "hoje", "último", "pessoas", "pessoa", "vez", "vezes", "apenas", "deve", "devem", "enquanto", "sido", "duas", "havia", "diz", "antes", "além", "segunda", "terça", "quarta", "quinta", "sexta", "feira", "cada", "vários", "várias", "domingo", "sexta-feira", "terça-feira", "segunda-feira", "alguns", "algumas", "quinta-feira", "quarta-feira", "sábado", "fazer", "porque", "sob", "têm","s", "v", "aqui", "então", "exª", "srª", "v.exª", "srs", "nº", "assim", "nesse", "sendo", "desse", "desa", "portanto", "aí", "art", "coisa", "qualquer", "quanto", "dessa", "srªs", "sras", "sr", "lá", "senhor", "todas", "tão", "nessa", "senhores", "disso", "alguma", "pois", "desses", "tendo", "sobretudo", "quais".

[40] There are more speeches in the Senate's Committee on Foreign Relations (CRE). Out of the total amount of speeches ($N = 62,410$), 44,005 speeches were made in the Senate's Committee, and 18,405 in the Chamber's Committee (CREDN). In relative terms, 70.5% of the database is composed by Senate's Committee speeches and 29.5% by Chamber's Committee speeches.

methods to assign the topics, as suggested by Grimmer and Stewart (2013, p.15).

Besides, LDA goes beyond grouping words into a set of topics. This model assumes that each document is a mixture of k underlying latent topics, and each topic has a certain distribution of words (Blei et al., 2003, p.996). Regarding the outputs, the LDA model gives a list of terms that contribute to each topic. Based on these terms, the researcher can create a label to each topic. The other outcome of the model is data on how much, in terms of percentage, each topic exists in each document. In sum, LDA calculates the influence of each word for each topic, and the importance of each topic per document. In the next section, I describe the steps to find the optimal number of k topics, the only input that the researcher has to insert in the model besides the actual documents being analyzed.

4.4.3 Finding the best k

The next step was to find out the best number of topics to put in the LDA model. As mentioned above, the information on the number of topics (k) is the only data that the researcher has to insert in the LDA model.

To come up with this number, I could have used the number of topics that the literature suggests. However, given that no previous study had analyzed these corpora, that was not an option.

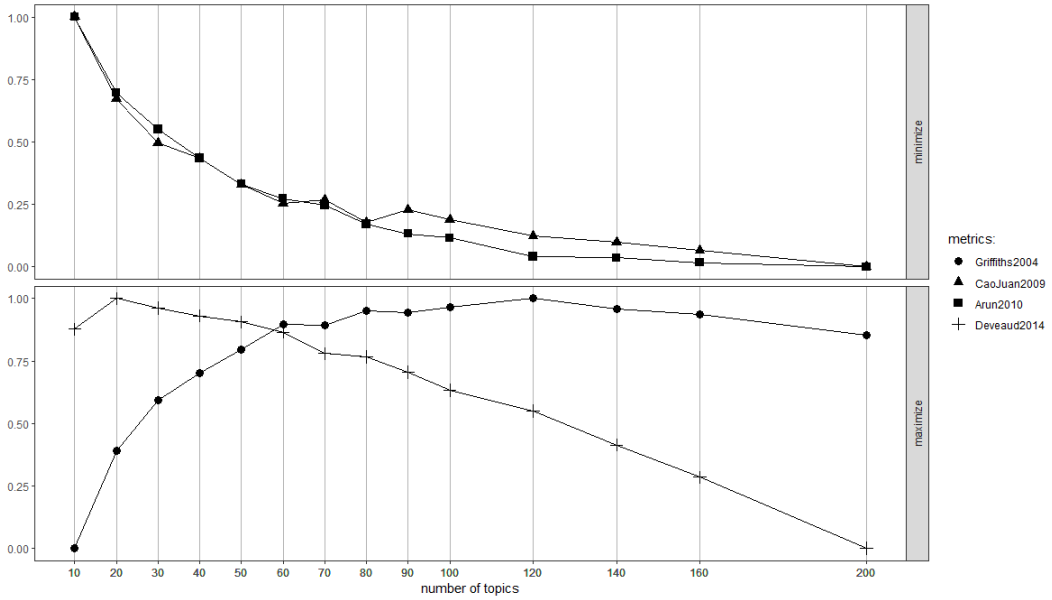
Moreover, because of the lack of information on the number of topics about foreign issues inside the Brazilian Congress, I opted to use a method that finds the optimal number of topics. The *R* package *ldatuning* has a command (`FindTopicsNumber`) that performs four different metrics that, together, can be used to justify the number of topics chosen. This functionality is compute-intense, and I ran it on the Congress's *dtm* varying the number of topics between: 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 120, 140, 160, 200.[41]

The resulting plot is shown in Figure 19.

According to Figure 19, the optimal number of topics is between 50 and 70. Despite the two metrics to be minimized, *Arun2010* (Arun et al., 2010) and *CaoJuan2009* (Cao et al., 2009), not being at their minimum values, which only occurs with $k = 200$, when we put together the statistics that should be maximized, *Griffiths2004* (Griffiths and Steyvers, 2004) and *Deveaud2014* (Deveaud et al., 2014), they are at their maximum

[41] The computer took 24 hours with parallel processing to run this command.

Figure 19: Number of topics



value. Additionally, the gain in reduction in the *Arun2010* and *CaoJuan2009* statistics is practically marginal after $k = 60$.

To get a better understanding on what the optimal k would be, I also ran a cross-validation function. I used the package *topicmodels* of R and its `LDA` function to generate the graph of 20,[42] and chose the model with the k that produces the lowest estimate of perplexity. This is because perplexity is a measure of how much a probability distribution can predict a sample based on the model created using the training data set.

Perplexity, then, is the difference between the results of the model applied to the training set and the results of the model applied to the test set. The bigger the difference, the worse the model. Thus, we need to choose a k with a low perplexity in Figure 20.

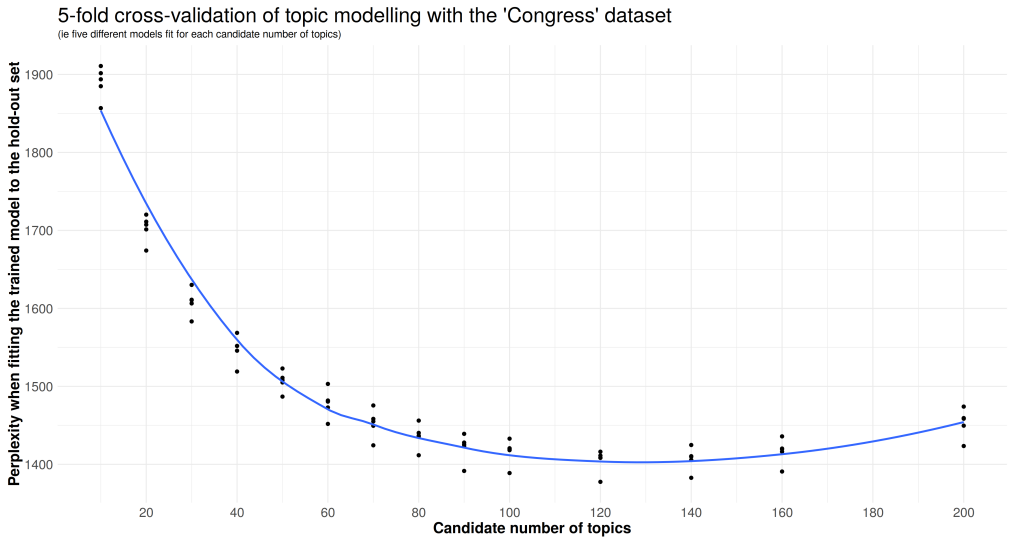
Figure 20 is a graphic representation of five validation tests (represented by the black dots in the graph) for each candidate k, the blue line being the line of the mean values of these tests. In addition, as in the test in Figure 19, I ran the cross-validation test for the following candidate *ks*: 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 120, 140, 160, 200. Finally, Figure 20 indicates that from $k = 60$ on the drop in perplexity is small, if we as compared with the rapid increase of k.

Therefore, I chose to run the LDA models with $k = 60$, considering the results of `FindTopicsNumber` command in conjunction with the cross-validation results.[43] Moreover,

[42] Construction of Figure 20 followed the instructions in the manual prepared by Nidhi: `http://www.rpubs.com/MNidhi/NumberoftopicsLDA`

[43] Although I chose to run the topic modeling with $k = 60$, I also run with k equal to 20, 40, and 50. In

Figure 20: Number of topics: Cross-validation



I ran the LDA model with Gibbs sampling, and I set the *burn-in* parameter to 100; the *iteration* parameter to 1000; and a *keep* parameter to 50. That means that the first 100 iterations were discarded, and that the model kept the log-likelihood of every 50 iterations for the subsequent 1,000 iterations. Finally, in relation to assigning topics to each document, I chose the topic that impacted each document the most. Below, I present the results of the topic model.

4.5 Results: Brazilian Congress' Committees on Foreign Relations

The results of topic modeling in the corpora of speeches made at the two Committees on Foreign Relations of the National Congress were very interesting. As we will see below, results show that Congress members are neither silent nor indifferent to the debate on international affairs. Besides proving the interest of Congress members on international issues, the results of the topic model have internal and external validity.

Since the analysis method was unsupervised, that is, we do not have preexisting comparison parameters, before analyzing the results it is necessary to verify if they are reliable and valid. For this, we have to verify the reliability, the internal validity and the external validity of the topics generated.

Regarding reliability, *id est*, the ability to obtain the same results by following the steps of a process, the LDA method may not be very reliable (Maier et al., 2018, p.99). This point is important, since, as it is a quantitative method, with a step-by-step script

Appendix B, I listed the results for those *ks* and labeled each of the categories generated.

of the process carried out, it is to be expected that other researchers will obtain the same results.

There is, however, a random component in the LDA algorithm, which is the point at which the algorithm begins to read the database. Depending on where the algorithm starts reading, results may change. To circumvent this reliability problem, I've run all models with two predetermined start points: $seed = 77$ and $seed = 1234$. This precaution ensures that results will always be the same when running the LDA model with the parameters used by me.[44]

As for the frequency with which each topic occurs, Figure 21 lists in decreasing order the proportion of speeches for each topic. The most frequent topic, with almost 6%, is the topic without identification (*NA*). Non-identification occurs for two reasons.[45]

On the one hand, when there are not enough words in the speech for the topic model algorithm to be able to assign the document (herein, a document is a speech) to a specific topic. On the other hand, the original speech has enough words, but, after cleaning verbs, adverbs and stop words, the "clean" speech does not have enough words for the model to assign it to a topic.

The following sentences are examples of these two situations: "Seguramente." [Surely], "Pois não" [No problem], and "Obrigado" [Thank you]. The word "seguramente" is an adverb, so it was excluded from the topic modeling, while the words "pois," "não" and "obrigado" are considered stop words, and have consequently been removed from the analysis.

After the *NA* category, the following five topics (Senator 1, Representative, Senator 3, Senator 2, Senate Committee) are all connected to treatment among senators and federal deputies. This finding was expected, since it is quite common for Congress members to refer to themselves or their colleagues, as well as to the procedures of the Committees, during their speeches. Thus, words like "senator", "deputy" and "report" appear very often.

The next topic is the first one to have a substantive meaning, *Ambassador appointment*, under which 3% of the speeches were classified. Among the words that have the most impact on the differentiation of this topic, we can find: "ambassador", "exhibition",

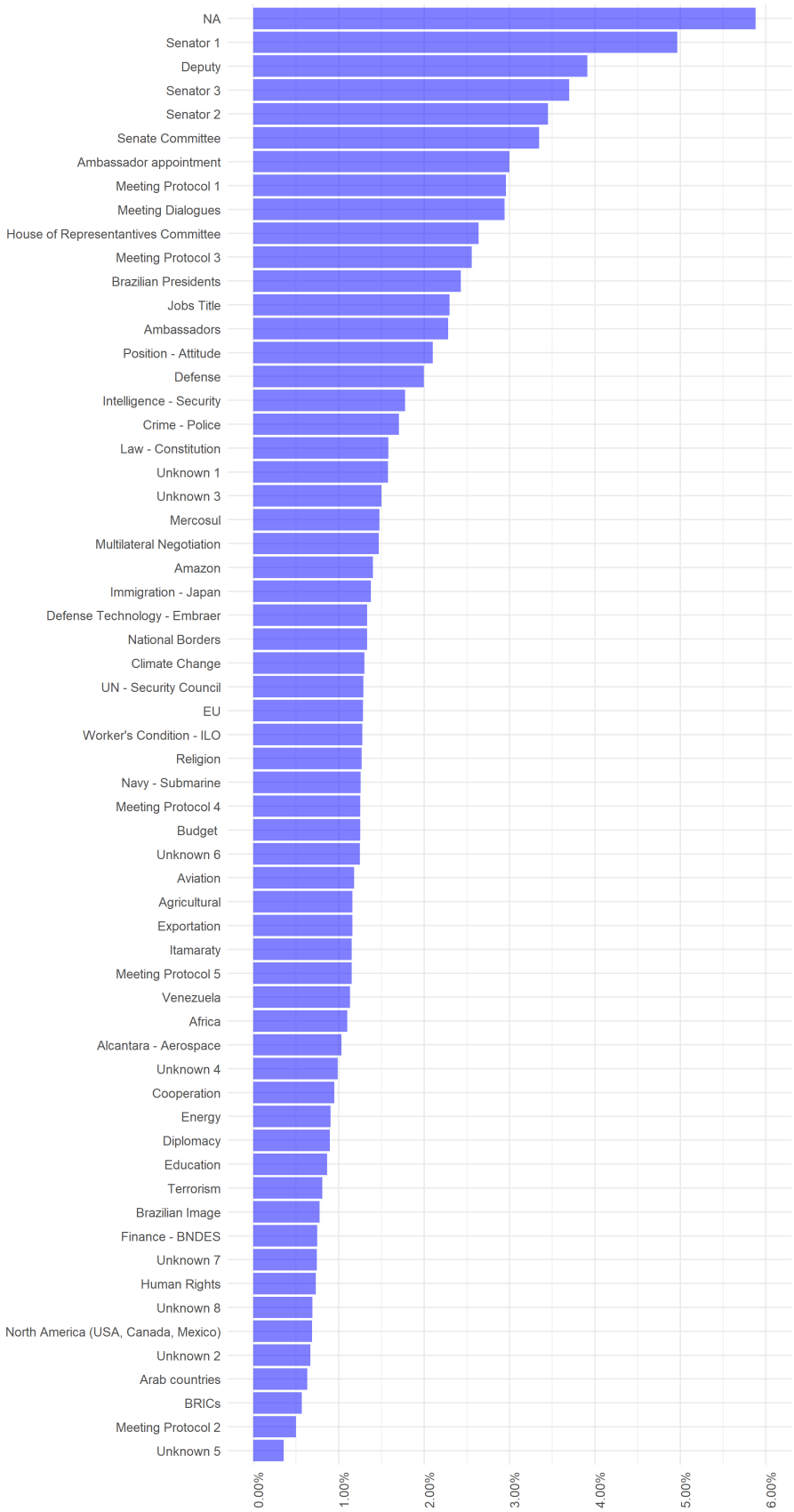
[44] There is a debate in the literature about the reliability of the LDA algorithm with random initialization. While most researchers state that the results are different if different *seeds* are used, DiMaggio et al. (2013) and Levy and Franklin (2014) argue that the results will be the same even with different *seeds*.

[45] When the LDA model created a topic, but the words that most influence it were not coherent, that is, there was no semantic relation between them, I coded the topic as "unknown". Out of the 60 topics of the LDA model chosen, 8 were classified as "unknown".

"appointed" and "mission".[46] These words are closely linked to the process of appointing an ambassador to a diplomatic mission abroad. Figure 21 below bears a complete list of the frequency with which all topics occur:

[46] As we shall see below, these results indicate the internal validity of the model.

Figure 21: Proportion of speeches about each topic ($k = 60$) throughout the period analyzed



Proceeding to validity, I will first discuss internal validity, which is the consistency of the results obtained. To see if the topics are consistent, I've created Table 16 with all 60 topics and the 30 words that most contribute to the creation of such topics.

As I mentioned in the methodology section, the LDA method assumes that all words influence, to a greater or lesser extent, all topics. In addition, the same document (or speech, in our database) may contain more than one topic. To make it easier to see the influence of each term/word on each topic, Table 16, Appendix B.4 lists the main words in order of impact on each topic. For example, the topic "Intelligence – Security" is more impacted by the word "information" then by the term "intelligence", "data", "system", "security", and so on.

In Table 16 we can verify that there is internal validity in the results obtained in the topic model with $k = 60$, since the words that most impact each topic have a semantic relation between them. For instance, the words that most impact topic 60 ("Venezuela") are, among others, "venezuela", "democracy", "venezuelan", "chávez" and "dictatorship".

Another example of internal consistency is provided by topic 50 ("Aviation"), where the strongest words are "air," "airplane," "varig," "passenger" and "infraero." In both topics mentioned the words that proportionately impact them the most have a correlation in their meanings or appear together when those themes are debated.

Regarding the debate on Venezuela, issues such as the political situation in the country and whether there was a break in the democratic principle in the country's elections are discussion points that appear routinely. On the subject of aviation, in addition to words that carry competing meanings, such as "airplane" and "air", we have a name of a Brazilian airline that filed for bankruptcy, Varig, and the name of the federal public company responsible for airports and air traffic in Brazil, Infraero. Therefore, there is internal validity in the topics obtained with LDA and $k = 60$.

The next step is to check for external validity in the results found, that is, assessing whether the topics generated have the ability to be generalized. To test external validity, I chose 4 substantive topics (as opposed to procedural topics such as "Senator 1" or "Representative") and created time series graphs of the proportion of speeches related to each of the following topics: Climate Change, Multilateral Negotiations, Ambassador Appointments and Venezuela.

As we will see below, it is possible to observe that these four international topics in the

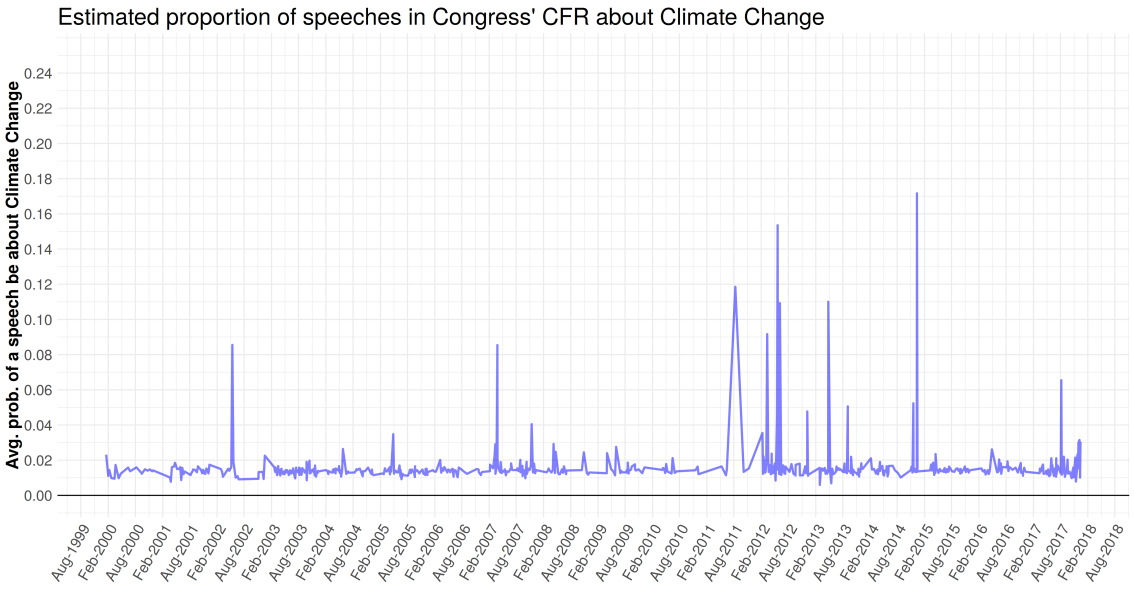
two Committees, CRE and CREDN, have a temporal dynamics. That is, the proportion of speeches on each theme varies over time, but not in a random way. The number of speeches on each theme is affected by events or situations that are intrinsically linked to the themes. This fact is fundamental, for it reveals that the model is able to capture time variations, and that the dynamics of speeches is also affected by events.

Figure 22 shows the evolution of speeches on the theme of climate change between January 2000 and December 2017 in the two Committees on Foreign Relations, CRE and CREDN. Throughout this period, we can see that there were some peaks with higher occurrence of speeches labeled with the theme of "climate change".

These peaks coincide with events external to the Committees that are directly connected with the environmental theme. For example, the first frequency peak of this topic occurred between May and June 2002, a period in which almost 8% of speeches in the two Committees regarded "climate change," according to the topic model. Those speeches preceded an important international meeting on environment, known as *Earth Summit 2002*, or *Rio+10*, which took place between August and September 2002 in the city of Johannesburg, South Africa.

Another period where we can observe external validity is the year 2012, when there were several peaks of greater frequency of speeches on climate change. That year, the United Nations Conference on Sustainable Development (UNCSD), Rio+20, took place in Rio de Janeiro. Brazil played a major role in these two events, since they were celebrating the 10-year and 20-year anniversaries, respectively, of ECO-92 (*Earth Summit*), held in the Brazilian city of Rio de Janeiro. Additionally, the greater number of speeches in 2012 shows that the model captured the topic well, since the members of the two Committees were expected to speak more about the theme in the year when an important international environmental event was hosted by Brazil.

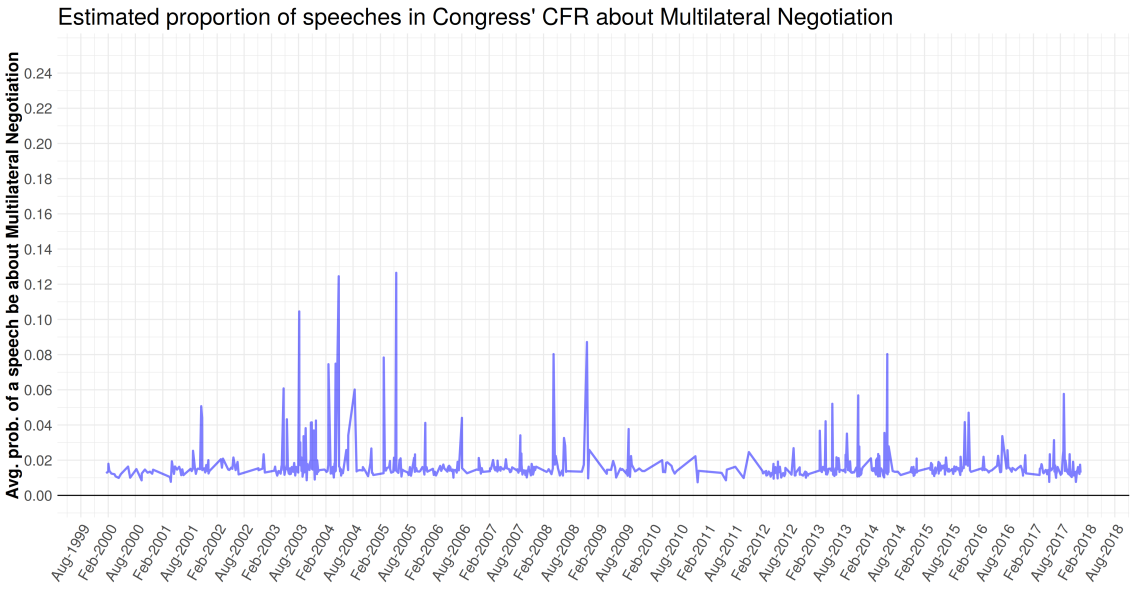
Figure 22: Congress CFR speeches about Climate Change



Another topic that shows that the LDA model used has external validity is "Multilateral Negotiations." Figure 23 shows that during the Lula Administration (2003–2011), the topic of multilateral negotiations appeared much more frequently in the two National Congress' Committees on Foreign Relations. This increased frequency occurred mainly during President Lula's first term. This fact is corroborated by the literature that shows that one of Lula administration's central axes in foreign policy was the Brazilian participation in multilateral forums, such as WTO and UN.

According to (de Almeida, 2004, p.166), "[while] the administration of Fernando Henrique Cardoso was characterized by moderate multilateralism (...), [t]he government of President Luiz Inácio Lula da Silva has a strong multilateralism and defends the sovereignty and equality of all countries with greater rhetorical emphasis than had been the case in the previous administration". This emphasis on multilateralism was quenched during the administrations of Dilma Rousseff and Michel Temer, a process that can also be observed in Figure 23.

Figure 23: Congress CFR speeches about Multilateral Negotiations

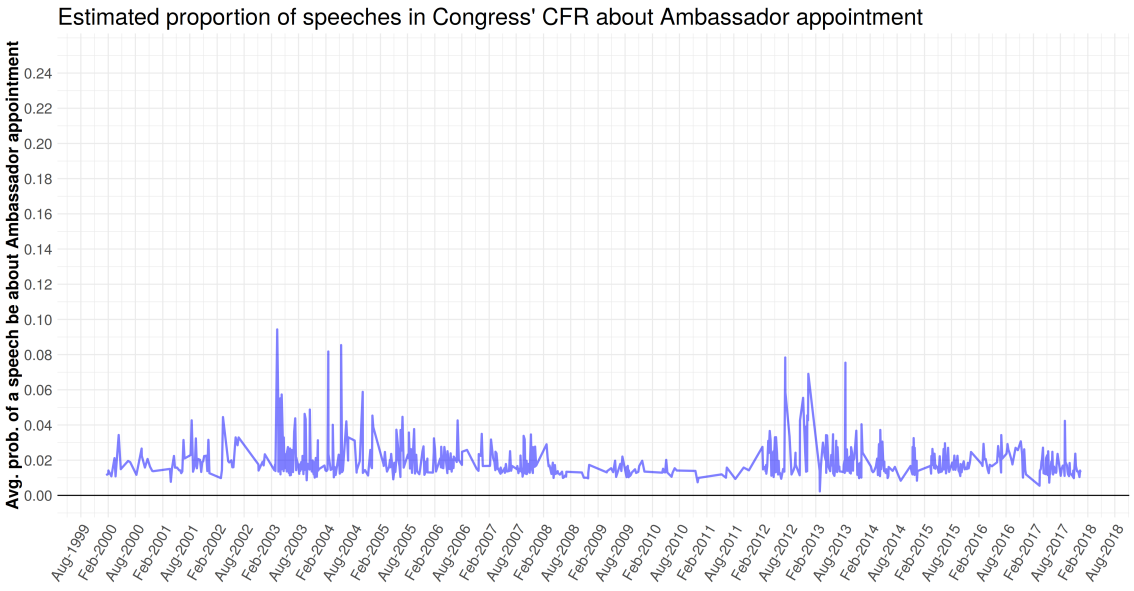


During the Lula administration, there were also proportionally more speeches on the topic "Ambassador Appointments", as we can see in Figure 24. In terms of external validity, this period was marked by a strong expansion in the number of Brazilian representations abroad. In the Lula Government, "52 embassies, 6 missions to International Organizations, 22 consulates and one diplomatic office, in Palestine" were inaugurated (Amorim, 2010, p.226).

In order to keep these new embassies and missions functioning, the diplomatic corps also had to be expanded, jumping from 1,000 diplomats in 2005 to 1,400 in 2010 (Amorim, 2010). Therefore, a bigger number of speeches regarding the appointment of ambassadors – an exclusive prerogative of CRE (Senate) – was expected to occur during the years of the Lula administration.[47]

[47] As CRE and CREDN data are grouped in Figure 24, the proportion of speeches on the appointment of ambassadors tends to be much higher in CRE than the almost 10% to which it arrives in March 2003.

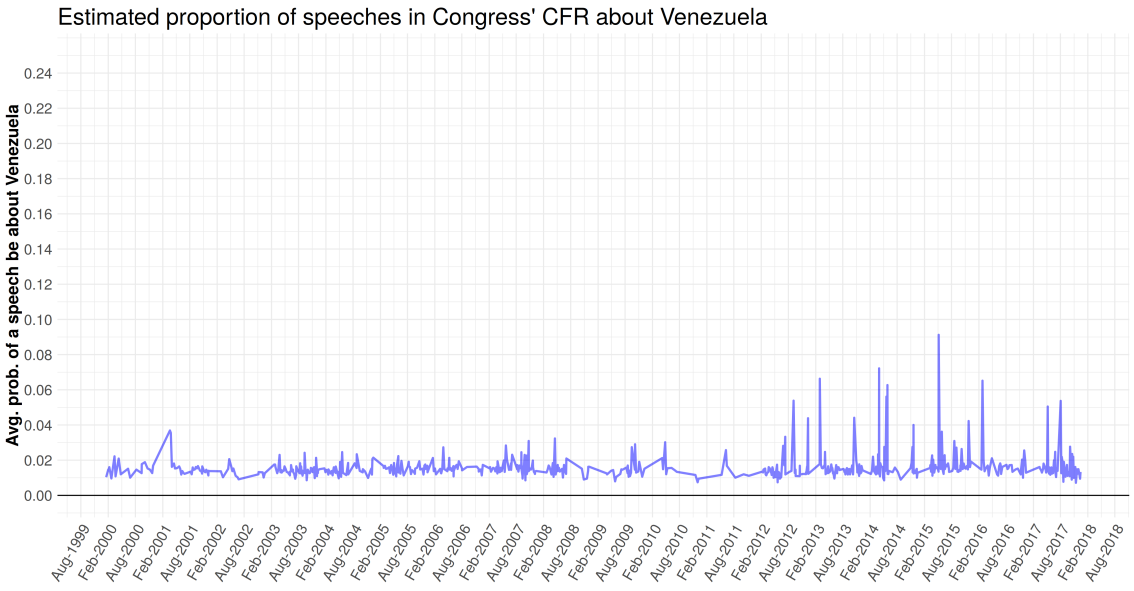
Figure 24: Congress CFR speeches about Ambassador Appointments



Finally, the "Venezuela" topic shows that the members of CRE and CREDN are not indifferent to the regional context. Figure 25 illustrates the proportion of speeches about Venezuela and shows a clear increase after Nicolas Maduro's rise to power in early 2013. After Maduro assumed the presidency, the economic situation in Venezuela has worsened a lot. The poverty level rose from 27% of the population in 2013 to 73% in 2015 (Merke et al., 2016, p.2).

Thus, during this most recent period of history, there has been a growing debate in both the media and academia on the negative consequences of the authoritarian regime to the Venezuelan population (van der Vaart et al., 2009; Merke et al., 2016). In this sense, it is interesting to note that, precisely in this period, the members and guests of CRE and CREDN have started to debate more about Venezuela. Therefore, Figure 25 indicates that the method used in this research has external validity.

Figure 25: Congress CFR speeches about Venezuela



In summary, the tables and figures referred to in this section demonstrate that the LDA method is reliable and has internal and external validity. Consequently, it is possible to use topic models to further examine the corpora of speeches given in the Committees on Foreign Relations of the National Congress of Brazil. This finding is of great importance, since it paves the way for the discourse analysis of a great historical period, 18 years, in a quantitative and unsupervised way. In the next section, I will discuss the main findings and point out a future research agenda, aiming to explore this new database with an innovative methodology.

4.6 Conclusion

Brazilian Congress members are neither indifferent nor absent from the debate on international issues. The quantitative analysis of the speeches of members and guests of the two External Relations Committees of the Brazilian Congress, CRE and CREDN, shed light on an area that until today has been practically unnoticed by Political Science and International Relations in Brazil, that is, the standing committees of the Congress.

However, this lack of attention was not due to the unimportance of such an analysis, but rather to the technical difficulty and the number of hours needed to analyze, read and categorize this vast material. In the two Congress' Committees on Foreign Relations alone, 62,410 oral statements were made between 2000 and 2017. Analyzing all this

material manually would be virtually impossible for a person or even a research group.

In order to overcome this difficulty, the analyses made so far either focused on the plenary sessions of the two Legislative Houses or studied the Committees through the lens of voting behavior of politicians rather than discourse analysis. It is in this sense that this paper aims to bring some contributions to the debate on international relations in Brazil.

First, the collection and systematization of a speech database was a pioneering work in the scope of the standing Committees. All speeches were separated, and their authors, identified. The identification of authors followed two steps: 1) verifying if the author of the speech was a politician or not; if they were, the script searched to which political party they belonged, and 2) in cases of non-politicians, the algorithm searched for the profession or bureaucratic position of the guest speaker. These steps made it possible to create an unprecedented database.

Second, once the database of speeches in the Committees had been cleaned and structured, it was possible to apply the models of quantitative analysis of text. This step required building a new structure, with the purpose of preparing the data for insertion into the LDA model. The only element that the researcher has to insert in this type of model is the number of topics (k) in the corpora. The tests carried out to assess perplexity, and others like *Arun2010*, *CaoJuan2009*, *Griffiths2004* and *Deveaud2014*, indicated that there are 60 different topics in the corpora of the two Committees on Foreign Relations of the National Congress of Brazil. Thus, this research showed that it is possible to work with a huge base, with more than 60 thousand oral statements, with size ranging between 1 and 11,000 words, in a personal computer. It further illustrates the possibility of performing quantitative research without the need for a cloud-computing infrastructure that only a few universities in the world have.

Last but not least, I have shown that the results obtained with the topic template in the Committees' corpora are reliable and have both internal and external validity. In terms of reliability, any researcher, by following the scripts executed in this research should get the same results presented herein. As for internal validity, the words and expressions that most influence the 60 topics obtained with the topic model are words that have similar semantic meanings or are words and expressions that appear within the debate of that topic. In relation to external validity, the probability of occurrence of the topics varies according to factors that are external to the Committees, but related to the themes. This

fact suggests a temporal dynamics in the frequency of speeches depending on the subject analyzed.

Once these steps are accomplished, my next objective is to gain understanding on two points. First, to find out what are the subjects discussed by politicians from different political parties. That way, I will be able to understand if there is a party-related difference between the 60 topics listed by the LDA model. Second, to analyze whether there is a difference in the frequency of topics when comparing politicians from the ruling party with those from opposition parties. This point will hopefully contribute to a whole literature of Brazilian political science, which argues that much of the Congress dynamics can be explained by the government-opposition clash.

5 Conclusion

This dissertation sought to understand what international issues Brazilian newspapers and politicians talk about, and when they talk about them. In doing so, it has shed light in two new corpora. First, a collection of news articles published in the two main Brazilian newspapers: *Folha de S. Paulo* and *O Estado de S. Paulo*. Second, a collection of political speeches given in the Brazilian Congress' Committees on Foreign Relations.

More specifically, I analyzed the news published in the international sections of *Folha de S. Paulo* and *O Estado de S. Paulo* between January 2000 and December 2018, but found that much of the news was concentrated between 2007 and 2017. As for the political speeches, I gathered all the statements made at the meetings of the two Committees on Foreign Relations of the National Congress of Brazil, CRE and CREDN, between January 2000 and December 2017.

The creation of these two corpora and its structuring into databases is an unprecedented product in the area of International Relations in Brazil. The corpus of news on international issues has a total of 174,515 news articles, made up of 132,863 articles from Folha and 41,652 from Estadão. Meanwhile, the corpus of political speeches has a total of 62,410 oral manifestations held in the two Committees on Foreign Relations: 44,005 statements in CRE (Senate) and 18,405 in CREDN (Chamber of Deputies).

The analysis of this vast amount of data (236,925) was only possible thanks to the emergence of new quantitative text analysis techniques. These new methods became more accessible with the increasing data processing capacity of personal computers and the appearance of statistical packages and functions that treat text as data. As an example of this recent process, the version 1.0 (stable version for use) of the *R* package called *quanteda* – which is applied to the corpora in this dissertation – was released in 2018.

One of the achievements of these new techniques is the possibility of using algorithms for the classification of texts instead of hiring human coders specialized in the subject being analyzed. This allows the study of an ocean of textual data to be done outside the major academic centers, which have sufficient human and financial resources at their disposal. However, the use of statistical techniques requires a level of knowledge and practice of programming that, for the time being, is not widespread in human sciences. As an example, all the quantitative steps of this dissertation were written in R and amounted

to almost 9,000 lines of code, distributed in 32 different script files.

This codification enabled the automatic collection of all international news in Folha's online collection and in the NexisLexis database, in the case of Estadão. I also wrote a code in R to do web scrap the pages of the Chamber of Deputies and the Senate in order to collect the speeches within their respective Committees on Foreign Relations, CREDN and CRE. After collecting these two corpora of documents, a lot of effort was required to clean and process the textual data, which included verifying missing data and analyzing the database structure.

Once the databases were clean, and each variable contained only the information that should be contained by it, I applied a non-supervised topic model known as Latent Dirichlet Allocation (LDA) to the data set. This model was chosen because its assumptions fit the characteristics of the corpora. LDA assumes that each document can be composed of a mix of topics, and that each topic consists of a set of words. Moreover, the only information that the researcher has to input in the LDA model is the number of topics (k) in the corpus.

As the analysis of these two corpora is unprecedented and therefore there is no k value to be found in the literature, I used cross-validation techniques to verify the perplexity, besides examining the following metrics: *Arun2010* (Arun et al., 2010); *CaoJuan2009* (Cao et al., 2009); *Griffiths2004* (Griffiths and Steyvers, 2004) and *Deveaud2014* (Deveaud et al., 2014). The analysis of these parameters indicated that there are 60 topics in the corpus of political speeches and 80 topics in the corpus of international news. With the values of k set, I then applied the topic model to each corpus.

The findings show that topic modeling with LDA can be used to analyze foreign affairs topics in Brazil. LDA method creates coherent topics (internal validity), whose connection to real world events can be verified (external validity). These findings are important for showing that it is possible to analyze texts on international subjects in an unsupervised way, that is, without any previous classification. Additionally, the findings of this research show that the Brazilian newspapers focus on issues related to world powers, such as European Union, Russia and China, but mainly the United States. They also reveal that the newspapers' dynamics follow rather the logic of war journalism, in which news on conflicts and wars are more recurrent than themes related to peace and diplomacy (peace journalism).

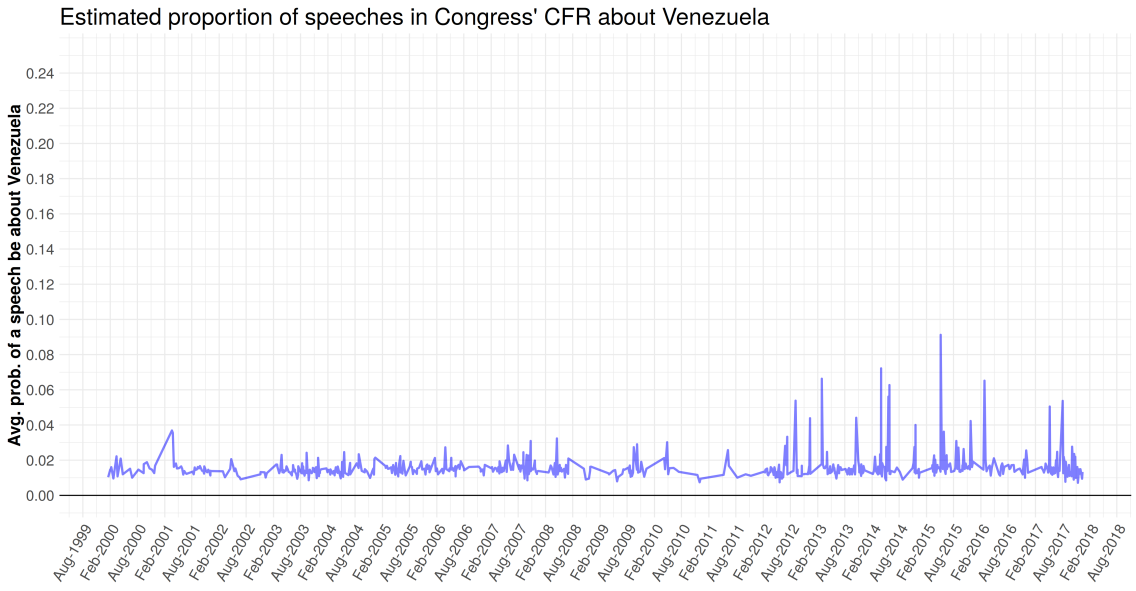
As for the topic model applied to the corpus of speeches made in the Committees on Foreign Relations, results show that the members of the Committees are neither silent nor indifferent to the debate on international affairs, since several topics that were relevant internationally were discussed in the two Committees. In fact, the proportion of these themes was in accordance with the occurrence of international events external to the Committees.

It is also relevant to note that the results do not clearly establish a connection between newspapers articles and political speeches on international issues in Brazil. The first major difference between the newspaper topics and the political speeches topics is that the former has a wider range of foreign affairs issues, with 80 different topics, while the politicians' speeches corpus has 60 different topics.

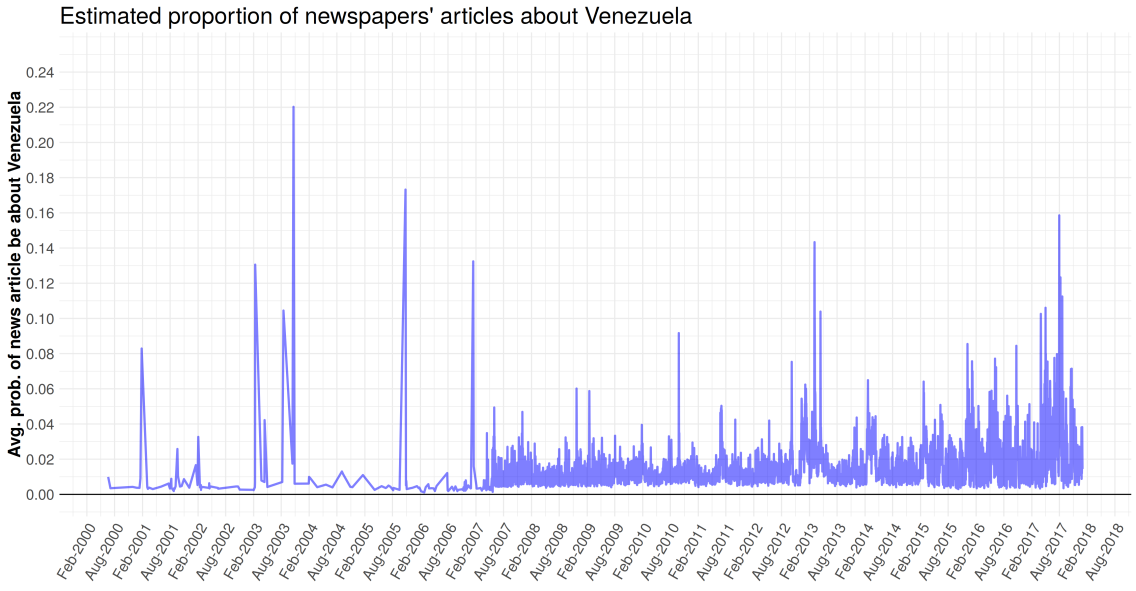
On one hand, Congressmen and Congresswomen in the Congress' Committees on Foreign Relations have a greater incentive to discuss topics in the agenda of each meeting, which means that they are more restricted than the journalists. On the other hand, journalists are influenced by events that occur all over the world, and the same journalist can write about different events/topics in the same day.

In terms of congruence between the topics from both corpus, there are twelve topics that occur in both models: European Union; Venezuela; Law–Constitution; UN–Security Council; Terrorism; Diplomacy; Immigration; Mercosur; Human Rights; Africa; Crime; and Intelligence. Nonetheless, their behavior over time do not match, except for the topic about Venezuela. Figure 26 reveals that both corpora experienced an increase in the number of either articles or political speeches about Venezuela after 2012.

Figure 26: News articles and politicians' speeches about Venezuela



(a) Congress' CFR speeches about Venezuela



(b) Proportion of newspapers articles about Venezuela

Besides showing that it is possible to use unsupervised topic modeling to analyze this type of corpus, this research opens up a promising future research agenda. Following are four paths that will be possible to pursue. First, in the next interaction of this research, I will focus on the differences and similarities between the two corpora, that is, news articles and politicians' speeches. Except for a few exceptions that try to analyze the Brazilian case (Azevedo, 2006; Figueiredo, 2015), the U.S. interpretation of the relationship between politics and the media prevails (Kurz, 1990; Robinson, 2001; Soroka, 2003b; Gentzkow and

Shapiro, 2010; Prat and Strömberg, 2011).

By attaining the goal of understanding the dynamics between mass media and political discourse, my research could help fill an important empirical gap. Elucidating this relationship will be an important step to explain the Brazilian case, but I also expect the results found to be expanded to other countries through comparative studies. This possibility is further widened by the very characteristic of the Legislative Branch of being more open to the media than the Executive:

> First, it has meant that the institutional power centers of Congress have been under the control of individuals who are quite willing to fight with the Executive. Second, it has meant that there are often several competing groups in the Congress, one of whom will probably be willing to form alliances with reporters. Third, younger reporters find they are of the same generation as the newer congressional leaders, especially subcommittee chairmen, and that they often have a common frame of reference. Next, the young members, especially subcommittee chairmen, actively seek out the press, recognizing that they need the press to build public opinion for proposals opposed by their more senior colleagues and the executive. (Kurz, 1990, p. 70)

This difference affects the relationship that mass media have with both the Executive and the Congress. Even though most of the foreign policy is designed and implemented by the Executive via Itamaraty, most of the meetings and decisions are taken behind closed doors. Furthermore, as a general rule Executive officers and diplomats do not give interviews, whereas Congressmen and Congresswomen are normally willing to talk and available for interviews in Congress.

This is relevant once the relationship between media and political discourse on international matters has achieved an unprecedented level of interaction around the world. Brexit and the U.S. presidential elections in the United States are an example of this puzzle. The former was partly characterized by a dispute between different newspapers[48] and how politicians used data on their speeches during the referendum campaign. The Sun, Telegraph, Express and Mail supported the Leave campaign, whereas BBC and The Guardian gave a rather lukewarm support for the Remain. Meanwhile, Boris Johnson, a

[48] http://www.royalgazette.com/opinion/article/20160713/brexit-not-shining-moment-for-british-media&template=mobileart

British politician, did a bus tour for the Vote Leave campaign.

In the case of the United States, there were rumors of Russia spreading fake news about the Democratic candidate, Hillary Clinton, thus smearing her image during the presidential campaign and ultimately impacting the election's results. All these facts highlight the importance of understanding how mass media and political discourse intertwine.

The second path for future a interaction of this research would be connecting the topic model analysis with roll-call votes (Gerrish and Blei, 2011). By doing so, we would be able to examine whether topic modeling can predict roll-call vote in Brazilian Committees on Foreign Affairs.

The third possible strategy would be running a dynamic model instead of a LDA model. Quinn et al. (2010) suggest that Dynamic Multitopic Model (DMM) is better for analyzing political speeches, since they usually have one topic per speech. Therefore, I will run a DMM in both corpora and check if the resulting topics are more coherent than the ones from LDA.

Finally, in the next phase of this research, I intend to analyze the words used by news outlets and by politicians. This next step will benefit from the growing literature on sentiment analysis on political texts and media slant (Gentzkow and Shapiro, 2010; Pak and Paroubek, 2010; Entman, 2010; Young and Soroka, 2012).

Regardless of the next steps that this research shall take, the findings presented herein will hopefully contribute to the advancement of empirical studies in the Brazilian literature on the relationship between foreign policy and media discourse. Moreover, the results of this research pave the way for better understanding how political preferences, especially those related to foreign policy, are expressed, and how such preferences impact media outlets whilst being impacted by them.

References

H. Allcott and M. Gentzkow. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–36, 2017.

G. Almond. *The American People and Foreign Policy.* Greenwood Press Reprint, Westport, Conn, 2 edition edition, Sept. 1977. ISBN 978-0-8371-9617-6.

G. A. Almond. The American people and foreign policy. 1950.

C. Amorim. Brazilian foreign policy under President Lula (2003-2010): An overview. *Revista brasileira de política internacional*, 53(SPE):214–240, 2010.

M. Arretche. Federalismo e democracia no Brasil: A visão da ciência política norte-americana. *São Paulo em Perspectiva*, 15(4):23–31, Dec. 2001. ISSN 0102-8839. doi: 10.1590/S0102-88392001000400004.

R. Arun, V. Suresh, C. V. Madhavan, and M. N. Murthy. On finding the natural number of topics with latent dirichlet allocation: Some observations. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 391–402. Springer, 2010.

C. Aurélio Pimenta de Faria. O Itamaraty e a Política Externa Brasileira: Do Insulamento à Busca de Coordenação dos Atores Governamentais e de Cooperação com os Agentes Societários. *Contexto internacional*, 34(1), 2012.

F. A. Azevedo. Mídia e democracia no Brasil: Relações entre o sistema de mídia e o sistema político. *Opinião Pública*, 12(1):88–113, 2006.

K. Benoit, K. Watanabe, H. Wang, P. Nulty, A. Obeng, S. Müller, and A. Matsuo. Quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, 30(3):774, 2018. doi: 10.21105/joss.00774.

A. J. Berinsky, G. A. Huber, and G. S. Lenz. Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk. *Political Analysis*, 20(03): 351–368, 2012. ISSN 1047-1987, 1476-4989. doi: 10.1093/pan/mpr057.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

D. J. Blood and P. C. Phillips. Recession headline news, consumer sentiment, the state of the economy and presidential popularity: A time series analysis 1989–1993. *International Journal of Public Opinion Research*, 7(1):2–22, 1995.

L. Boxell, M. Gentzkow, and J. M. Shapiro. Is the internet causing political polarization? Evidence from demographics. Technical report, National Bureau of Economic Research, 2017.

J. C. Brandi Aleixo. Fundamentos e Linhas Gerais da Politica Externa do Brasil. *-68 Revista Brasileira Estudos Politicos*, 67:7, 1988.

J. Cao, T. Xia, J. Li, Y. Zhang, and S. Tang. A density-based method for adaptive LDA model selection. *Neurocomputing*, 72(7-9):1775–1781, 2009.

J. A. Cheibub, A. Figueiredo, and F. Limongi. Partidos políticos e governadores como determinantes do comportamento legislativo na câmara dos deputados, 1988-2006. *Dados-Revista de Ciências Sociais*, 52(2), 2009.

A. L. R. da Silva. O Brasil diante da globalização: A política externa do governo Fernando Henrique Cardoso (1995-2002). *Carta Internacional*, 7(1):20–34, 2012.

R. S. da Silva. *A política externa brasileira analisada em três dimensões: um estudo sobre a comissão de relações exteriores e de defesa nacional da câmara dos deputados.* masterThesis, UFPE, Recife, Mar. 2012.

R. S. da Silva. *Os Parlamentares são Omissos ao Debate da Política Externa? Um Exame dos Atos Internacionais no Congresso Nacional.* PhD thesis, Universidade Federal de Pernambuco, Recife, 2016.

P. R. de Almeida. Uma política externa engajada: A diplomacia do governo Lula. *Revista Brasileira de Política Internacional*, 47(1):162–184, 2004.

R. Deveaud, E. SanJuan, and P. Bellot. Accurate and effective latent concept modeling for ad hoc information retrieval. *Document numérique*, 17(1):61–84, 2014.

P. DiMaggio, M. Nag, and D. Blei. Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of US government arts funding. *Poetics*, 41(6):570–606, 2013.

J. N. Druckman, D. P. Green, J. H. Kuklinski, and A. Lupia. *Cambridge Handbook of Experimental Political Science*. Cambridge University Press, Cambridge ; New York, 2011. ISBN 978-0-521-19212-5.

I. D. Duchacek. *Comparative Federalism: The Territorial Dimension of Politics, Modern Comparative Politics Series*. New York: Holt Rinehart and Winston, 1970.

A. C. Eggers and J. Hainmueller. MPs for Sale? Returns to Office in Postwar British Politics. *American Political Science Review*, 103(4):513–533, Nov. 2009. ISSN 1537-5943, 0003-0554. doi: 10.1017/S0003055409990190.

P. K. Enns. The public's increasing punitiveness and its influence on mass incarceration in the United States. *American Journal of Political Science*, 58(4):857–872, 2014.

R. M. Entman. Media framing biases and political power: Explaining slant in news of Campaign 2008 - Robert M. Entman, 2010. *Journalism*, 11(4):389–408, 2010.

M. S. Evans. A computational approach to qualitative analysis in large textual datasets. *PloS one*, 9(2):e87908, 2014.

R. R. Figueiredo. Mídia e eleições: Cobertura jornalística da campanha presidencial de 1994. *Opinião Pública*, 5(1):72–89, 2015.

P. C. D. Fonseca and S. M. M. Monteiro. Credibilidade e populismo no Brasil: A política econômica dos governos Vargas e Goulart. *Revista Brasileira de Economia*, 59(2):215–243, 2005.

B. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315(5814):972–976, 2007.

S. K. Gadarian. Foreign Policy at the Ballot Box: How Citizens Use Foreign Policy to Judge and Choose Candidates. *The Journal of Politics*, 72(04):1046–1062, Oct. 2010. ISSN 1468-2508. doi: 10.1017/S0022381610000526.

J. Galtung. Peace journalism. *Media Asia*, 30(3):177–180, 2003.

W. A. Gamson and A. Modigliani. Media discourse and public opinion on nuclear power: A constructionist approach. *American journal of sociology*, 95(1):1–37, 1989.

C. Gautrais, P. Cellier, R. Quiniou, and A. Termier. Topic Signatures in Political Campaign Speeches. In *EMNLP 2017-Conference on Empirical Methods in Natural Language Processing*, 2017.

M. Gentzkow. Polarization in 2016. *Toulouse Network for Information Technology Whitepaper*, 2016.

M. Gentzkow and J. M. Shapiro. What drives media slant? Evidence from US daily newspapers. *Econometrica*, 78(1):35–71, 2010.

S. Gerrish and D. M. Blei. Predicting legislative roll calls from text. In *Proceedings of the 28th International Conference on Machine Learning (Icml-11)*, pages 489–496, 2011.

M. Gilens. Why American Hate Welfare. *Race, Media, and the Politics of Antipoverty Policy, Chicago, Presses Universitaires de Chicago*, 1999.

D. Greene and J. P. Cross. Unveiling the Political Agenda of the European Parliament Plenary: A Topical Analysis. *arXiv:1505.07302 [cs]*, May 2015.

T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl 1):5228–5235, 2004.

J. Grimmer. A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases. *Political Analysis*, 18(01):1–35, 2010. ISSN 1047-1987, 1476-4989. doi: 10.1093/pan/mpp034.

J. Grimmer and G. King. General purpose computer-assisted clustering and conceptualization. *Proceedings of the National Academy of Sciences*, 108(7):2643–2650, 2011.

J. Grimmer and B. M. Stewart. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, 21(3):267–297, July 2013. ISSN 1047-1987. doi: 10.1093/pan/mps028.

A. Hamilton, J. Madison, J. Jay, and J. R. Pole. *The Federalist*. Hackett Publishing, 2005.

M. Haselmayer and M. Jenny. Measuring the tonality of negative campaigning: Combining a dictionary approach with crowd-coding. *Political context matters: Content analysis in the social sciences. University of Mannheim*, 2014.

D. Hillard, S. Purpura, and J. Wilkerson. Computer-Assisted Topic Classification for Mixed-Methods Social Science Research. *Journal of Information Technology & Politics*, 4(4):31–46, May 2008. ISSN 1933-1681. doi: 10.1080/19331680801975367.

O. R. Holsti. Public Opinion and Foreign Policy: Challenges to the Almond-Lippmann Consensus. *International Studies Quarterly*, 36(4):439–466, Dec. 1992. ISSN 0020-8833. doi: 10.2307/2600734.

O. R. Holsti. *Public Opinion and American Foreign Policy*. University of Michigan Press, Ann Arbor, revised edition edition, June 2004. ISBN 978-0-472-03011-8.

D. J. Hopkins and G. King. A Method of Automated Nonparametric Content Analysis for Social Science. *American Journal of Political Science*, 54(1):229–247, Jan. 2010a. ISSN 00925853, 15405907. doi: 10.1111/j.1540-5907.2009.00428.x.

D. J. Hopkins and G. King. A Method of Automated Nonparametric Content Analysis for Social Science. *American Journal of Political Science*, 54(1):229–247, Jan. 2010b. ISSN 00925853, 15405907. doi: 10.1111/j.1540-5907.2009.00428.x.

D. J. Hopkins, E. Kim, and S. Kim. Does newspaper coverage influence or reflect public perceptions of the economy? *Research & Politics*, 4(4):2053168017737900, Oct. 2017. ISSN 2053-1680. doi: 10.1177/2053168017737900.

S. Iyengar and D. R. Kinder. News that matters: Agenda-setting and priming in a television age. *News that Matters: Agenda-Setting and Priming in a Television Age*, 1987.

S. Jackman. Data from Web into R. *Political Methodologist*, 14(2), 2006.

C. Jacobi, W. van Atteveldt, and K. Welbers. Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital Journalism*, 4(1):89–106, Jan. 2016. ISSN 2167-0811. doi: 10.1080/21670811.2015.1093271.

A. K. Jain, M. N. Murty, and P. J. Flynn. Data Clustering: A Review. *ACM Comput. Surv.*, 31(3):264–323, Sept. 1999. ISSN 0360-0300. doi: 10.1145/331499.331504.

R. L. Jones and R. E. Carter. Some procedures for estimating" news hole" in content analysis. *Public Opinion Quarterly*, pages 399–403, 1959.

K. F. Kahn and P. J. Kenney. Do Negative Campaigns Mobilize or Suppress Turnout? Clarifying the Relationship between Negativity and Participation. *The American Political Science Review*, 93(4):877–889, Dec. 1999. ISSN 0003-0554. doi: 10.2307/2586118.

R. Keeble, J. Tulloch, and F. Zollman. *Peace Journalism, War and Conflict Resolution*. Peter Lang, 2010.

P. M. Kellstedt. Media Framing and the Dynamics of Racial Policy Preferences. *American Journal of Political Science*, 44(2):245–260, 2000. ISSN 0092-5853. doi: 10.2307/2669308.

R. J. Kurz. Congress and the Media: Forces in the Struggle Over Foreign Policy. In *The Media and Foreign Policy*, pages 65–78. Springer, 1990.

M. Laver and J. Garry. Estimating Policy Positions from Political Texts. *American Journal of Political Science*, 44(3):619–634, 2000. ISSN 0092-5853. doi: 10.2307/2669268.

S. T. Lee and C. C. Maslog. War or peace journalism? Asian newspaper coverage of conflicts. *Journal of Communication*, 55(2):311–329, 2005.

K. E. Levy and M. Franklin. Driving regulation: Using topic models to examine political contention in the US trucking industry. *Social Science Computer Review*, 32(2):182–194, 2014.

W. Lippmann. *Essays in the Public Philosophy*. Transaction Publishers, 1955.

W. Lippmann. *Public Opinion*. CreateSpace Independent Publishing Platform, Charleston, SC., Jan. 2010. ISBN 978-1-4505-3390-4.

T. Loughran and B. Mcdonald. When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance*, 66(1):35–65, Feb. 2011. ISSN 1540-6261. doi: 10.1111/j.1540-6261.2010.01625.x.

D. Maier, A. Waldherr, P. Miltner, G. Wiedemann, A. Niekler, A. Keinert, B. Pfetsch, G. Heyer, U. Reber, T. Häussler, H. Schmid-Petri, and S. Adam. Applying LDA Topic Modeling in Communication Research: Toward a Valid and Reliable Methodology. *Communication Methods and Measures*, 12(2-3):93–118, Apr. 2018. ISSN 1931-2458. doi: 10.1080/19312458.2018.1430754.

C. D. Manning, P. Raghavan, and H. Schutze. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, 2008. ISBN 978-0-521-86571-5.

J. H. Martin and D. Jurafsky. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Pearson/Prentice Hall Upper Saddle River, 2009.

C. Martindale. *Romantic Progression: The Psychology of Literary History*. Hemisphere Publishing Corporation, 1975.

Martindale, Colin. *The Clockwork Muse: The Predictability of Artistic Change*. Basic NewYork, 1990.

F. Merke, A. E. Feldmann, and O. della Costa Stuenkel. Venezuela on the Edge: Can the Region Help? *Carnegie Endowment for International Peace*, 2016.

A. Mintz. The feasibility of the use of samples in content analysis. *The Language of Politics: Studies in Quantitative Semantics*, pages 127–52, 1949.

B. L. Monroe, M. P. Colaresi, and K. M. Quinn. Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict. *Political Analysis*, 16 (04):372–403, 2008. ISSN 1047-1987, 1476-4989. doi: 10.1093/pan/mpn018.

C. S. Montesquieu. *The Spirit of the Laws*. T. Evans, 1777, London, 1748.

D. C. Moreira. *Com a palavra os nobres deputados: frequência e ênfase temática dos discursos dos parlamentares brasileiros*. PhD thesis, Universidade de São Paulo, 2016.

G. Müller. Comissões e partidos políticos na Câmara dos Deputados: Um estudo sobre os padrões partidários de recrutamento para as comissões permanentes. *DADOS–Revista de Ciências Sociais*, 48:371–394, 2005.

T. E. Nelson, R. A. Clawson, and Z. M. Oxley. Media framing of a civil liberties conflict and its effect on tolerance. *American Political Science Review*, 91(3):567–583, 1997.

D. Newman, C. Chemudugunta, P. Smyth, and M. Steyvers. Analyzing entities and topics in news articles using statistical topic models. In *International Conference on Intelligence and Security Informatics*, pages 93–104. Springer, 2006.

J. M. Nicolau. As distorções na representação dos estados na Câmara dos Deputados brasileira. *Dados*, 40(3), 1997.

A. Pak and P. Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, volume 10, 2010.

J. W. Pennebaker, M. E. Francis, and R. J. Booth. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001, 2001.

J. W. Pennebaker, C. K. Chung, M. Irel, A. Gonzales, and R. J. Booth. Linguistic inquiry and word count: LIWC 2007, 2007.

C. Pereira and B. Mueller. Uma teoria da preponderância do poder Executivo: O sistema de comissões no Legislativo brasileiro. *Revista Brasileira de Ciências Sociais*, 15(43): 45–67, 2000.

C. Pereira and B. Mueller. Comportamento Estratégico em Presidencialismode Coalizão: As Relações entre Executivo e Legislativo na Elaboração do Orçamento Brasileiro. *Dados*, 45(2):265–301, 2002. ISSN 0011-5258. doi: 10.1590/S0011-52582002000200004.

J. R. Petrocik. Issue Ownership in Presidential Elections, with a 1980 Case Study. *American Journal of Political Science*, 40(3):825–850, Aug. 1996. ISSN 0092-5853. doi: 10.2307/2111797.

C. A. Pimenta de Faria. Opinião pública e política externa: Insulamento, politização e reforma na produção da política exterior do Brasil. *Revista Brasileira de Política Internacional*, 51(2), 2008.

M. Porter. An algorithm for suffix stripping. Program 14(3):130–37. *Program*, 14(3): 130–37, 1980.

A. Prat and D. Strömberg. The political economy of mass media. *CEPR Discussion Paper No. DP8246*, 2011.

K. M. Quinn, B. L. Monroe, M. Colaresi, M. H. Crespin, and D. R. Radev. An automated method of topic-coding legislative speech over time with application to the 105th-108th US Senate. In *Midwest Political Science Association Meeting*, pages 1–61, 2006.

K. M. Quinn, B. L. Monroe, M. Colaresi, M. H. Crespin, and D. R. Radev. How to Analyze Political Attention with Minimal Assumptions and Costs. *American Journal of Political Science*, 54(1):209–228, Jan. 2010. ISSN 1540-5907. doi: 10.1111/j.1540-5907.2009.00427.x.

W. H. Riker. *Federalism: Origin, Operation, Significance*. Boston: Little, Brown, 1964.

T. Risse-Kappen. Public Opinion, Domestic Structure, and Foreign Policy in Liberal Democracies. *World Politics*, 43(4):479–512, July 1991. ISSN 1086-3338, 0043-8871. doi: 10.2307/2010534.

T. Risse-Kappen. *Bringing Transnational Relations Back In: Non-State Actors, Domestic Structures and International Institutions*. Cambridge University Press, Sept. 1995. ISBN 978-0-521-48441-1.

M. Roberts and M. McCombs. Agenda setting and political advertising: Origins of the news agenda. *Political communication*, 11(3):249–262, 1994.

P. Robinson. Theorizing the Influence of Media on World Politics: Models of Media Influence on Foreign Policy. *European Journal of Communication*, 16(4):523–544, 2001.

J. N. Rosenau. *Domestic Sources of Foreign Policy*. Free Press, New York, 1967.

S. T. Schmitt. *A Política Externa e o Poder Legislativo: Um Olhar Sobre a Comissão de Relações Exteriores e Defesa Nacional Do Senado Federal*. PhD Thesis, Universidade de São Paulo, 2011.

J. Sides. The Origins of Campaign Agendas. *British Journal of Political Science*, 36(03): 407–436, July 2006. ISSN 1469-2112. doi: 10.1017/S0007123406000226.

M. M. Soares and L. C. Lourenço. A representação política dos estados na federação brasileira. *Revista Brasileira de Ciências Sociais*, 19(56), 2006.

S. N. Soroka. Media, Public Opinion, and Foreign Policy. *Harvard International Journal of Press/Politics*, 8(1):27–48, Jan. 2003a. ISSN 1081-180X. doi: 10.1177/1081180X02238783.

S. N. Soroka. Media, public opinion, and foreign policy. *The International Journal of Press/Politics*, 8(1):27–48, 2003b.

G. H. Stempel. Sample Size for Classifying Subject Matter in Dailies. *Journalism and Mass Communication Quarterly*, 29(3):333, 1952.

B. M. Stewart and Y. M. Zhukov. Use of force and civil–military relations in Russia: An automated content analysis. *Small Wars & Insurgencies*, 20(2):319–343, June 2009. ISSN 0959-2318, 1743-9558. doi: 10.1080/09592310902975455.

P. J. Stone, D. C. Dunphy, M. S. Smith, and D. M. Ogilvie. *General Inquirer: A Computer Approach to Content Analysis*. The MIT Press, Dec. 1966. ISBN 978-0-262-69011-9.

G. d. S. Tarouco. Brazilian Parties According to their Manifestos: Political Identity and Programmatic Emphases. *Brazilian Political Science Review*, 5(1):54–76, 2011. ISSN 1981-3821.

G. d. S. Tarouco and R. M. Madeira. Esquerda e direita no sistema partidário brasileiro: análise de conteúdo de documentos programáticos. *Revista Debates*, 7(2):93–114, Aug. 2013. ISSN 1982-5269.

Y. R. Tausczik and J. W. Pennebaker. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology*, 29(1): 24–54, 2010.

A. W. van der Vaart, S. Dudoit, and M. J. van der Laan. Oracle inequalities for multi-fold cross validation. *Statistics & Decisions*, 24(3):351–371, 2009. doi: 10.1524/stnd.2006.24.3.351.

L. Young and S. Soroka. Affective News: The Automated Coding of Sentiment in Political Texts. *Political Communication*, 29(2):205–231, Apr. 2012. ISSN 1058-4609, 1091-7675. doi: 10.1080/10584609.2012.671234.

W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li. Comparing Twitter and Traditional Media Using Topic Models. In *Advances in Information Retrieval*, Lecture Notes in Computer Science, pages 338–349. Springer, Berlin, Heidelberg, Apr. 2011. ISBN 978-3-642-20160-8 978-3-642-20161-5. doi: 10.1007/978-3-642-20161-5_34.

C. Zirn and H. Stuckenschmidt. Multidimensional topic analysis in political texts. *Data & Knowledge Engineering*, 90:38–53, 2014.

A Appendix: Newspaper

Table 8: Newspapers: Database summary

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
newspaper	174,515	Categorical	–	–	–	–	–
news_date	174,515	Date	–	–	–	–	–
news_author	167,056	Categorical	–	–	–	–	–
news_title	174,515	Categorical	–	–	–	–	–
news_text	174,515	Categorical	–	–	–	–	–
news_section	174,515	Categorical	–	–	–	–	–
news_topic	41,652	Categorical	–	–	–	–	–
news_length_wd	174,515	350.9	260.7	1	192	448.5	10,426
news_length_char	174,515	2,210.6	1,628.7	6	1,220	2,829	64,841
news_load_date	41,650	Date	–	–	–	–	–

A.1 Terms when $k = 40$

Table 9: Newspaper Topics ($k = 40$)

Topic		Keys (Top 30)
1	**Latin America**	brasil, presidente, brasileiro, país, zelaya, lula, honduras, dilma, embaixada, relação, político, crise, paraguai, golpe, embaixador, chanceler, interino, oea, presidência, josé, mercosul, lugo, uruguai, itamaraty, micheletti, luiz, hondurenho, reunião, internacional, brasília
2	**Law - Constitution**	direito, lei, decisão, humano, projeto, medido, liberdade, comissão, contra, congresso, novo, texto, público, aprovado, nacional, medida, cidadão, legal, câmara, aprovação, constituição, senado, legislação, favor, supremo, proibição, sistema, constitucional, deputado, maioria
3	**UN - Security Council**	onu, unido, internacional, conselho, segurança, nação, país, organização, reunião, estado, resolução, relação, membro, exterior, paz, representante, missão, secretário-geral, comunidade, embaixador, situação, declaração, diálogo, negociação, diplomata, ban, humanitário, solução, assunto, conferência

4	**Unknown 1**	político, mundo, tempo, guerra, grande, fato, próprio, história, tradução, claro, problema, maneiro, questão, verdade, difícil, lugar, sociedade, exemplo, democracia, coisa, momento, razão, bom, lado, único, ano, diferente, importante, nenhum, pontar
5	**News**	jornal, jornalista, the, vídeo, tv, redar, imagem, site, times, mídia, canal, foto, internet, reportagem, notícia, mensagem, new, york, news, repórter, emissor, comunicação, twitter, página, publicado, of, informação, televisão, publicação, facebook
6	**Election**	eleição, eleitoral, voto, político, presidente, votação, presidencial, oposição, campanha, vitória, partido, maioria, eleitor, nacional, resultado, mandatar, candidato, pesquisa, novo, urna, líder, turno, coalizão, eleito, deputado, legislativo, conservador, cadeira, popular, pleito
7	**USA Election**	republicano, obama, democrata, campanha, trump, hillary, senador, mccain, clinton, eleitor, romney, branco, estado, barack, john, primário, presidente, governador, presidencial, presidência, senado, donald, delegado, político, convenção, flórido, contra, eleição, york, pesquisa
8	**Judicial System - Crime**	prisão, crime, tribunal, acusação, ano, contra, morte, juiz, julgamento, preso, investigação, assassinato, acusado, detenção, advogado, promotor, condenado, sentença, autoridade, prisioneiro, judicial, decisão, defeso, audiência, direito, condenação, denúncia, vítima, promotoria, ordem
9	**Economy - IMF**	país, banco, crise, economia, financeiro, euro, bilião, dívida, europeu, medida, zonar, público, grécia, central, crescimento, econômica, fiscal, grego, novo, fmi, título, pib, pacote, monetário, gasto, maior, finanças, prazo, internacional, mercado

10	**Russia - Argentina**	russo, rússia, argentino, presidente, putin, cristina, ucrânia, kirchner, país, aires, buenos, ucraniano, região, vladimir, separatista, soviético, geórgia, medvedev, macri, união, kremlin, província, líder, kiev, clarín, ex-presidente, crimeia, néstor, político, último
11	**Syria**	sírio, regime, assad, país, rebelde, contra, bashar, cidade, damasco, oposição, conflito, civil, arma, líbano, árabe, grupo, guerra, libanês, humano, violência, ditador, morto, força, direito, al, onu, mês, opositor, repressão, química
12	**USA**	americano, eua, obama, estado, presidente, unido, washington, barack, branco, secretário, bush, departamento, george, americana, novo, relação, w, congresso, guerra, john, clinton, kerry, hillary, viagem, administração, defeso, norte-americano, robert, pentágono, setembro
13	**China - India - Bolivia**	chinar, chinês, índio, pequim, país, morales, indiano, ano, paz, bolívia, região, oficial, líder, boliviano, prêmio, comunista, autoridade, nobel, evo, novo, agência, hong, província, kong, tibetano, relação, xinhua, asiático, taiwan, estatal
14	**Aviation**	avião, aeroporto, aéreo, passageiro, companhia, autoridade, voo, aeronave, trem, furacão, brasília, novo, local, km, hora, tempestade, passagem, horário, noite, porta-voz, agência, h, airlines, área, york, quilômetros, serviço, manhã, direção, próximo
15	**Global Health - WHO**	médico, hospital, caso, morte, doença, país, gripar, tratamento, vírus, novo, número, organização, suíno, paciente, autoridade, mundial, confirmado, ministério, oms, contra, dor, problema, estado, cincar, exame, unido, comum, sintoma, especialista, epidemiar

16	**Afghanistan - Pakistan - Terrorism**	afeganistão, paquistão, afegão, país, taleban, paquistanês, contra, bin, otan, laden, província, segurança, morte, terrorista, insurgente, força, tropa, morto, militante, líder, operação, porta-voz, karzai, região, civil, talebans, internacional, fonte, cabul, chefe
17	**Mexico - Haiti - Drugs - Immigration**	país, méxico, imigrante, haiti, refugiado, mexicano, fronteirar, droga, ilegal, haitiano, número, autoridade, cidade, imigração, milhão, tráfico, capital, príncipe, maconha, unido, violência, cartel, organização, morto, corpo, nigéria, ano, maioria, criminoso, estrangeiro
18	**Catholic Church**	igreja, católico, religioso, vaticano, deus, francisco, bento, cristão, santo, mundo, paulo, ano, padrar, cardeal, fé, fiel, religião, novo, homem, joão, pontífice, cerimônia, grande, pedrar, abuso, 2º, missar, comunidade, amor, palavra
19	**Natural Disasters**	terremoto, região, água, cidade, casa, área, tremor, local, forte, dano, quilômetros, autoridade, morto, chuva, grau, km, vítima, morador, emergência, agência, província, tsunami, magnitude, desastre, japão, incêndio, número, terra, sul, serviço
20	**Intelligence - Spy - Wikileaks**	informação, serviço, inteligência, documento, segurança, agência, funcionário, dado, investigação, site, agente, secreto, espionagem, autoridade, internet, wikileaks, acesso, nacional, relatório, departamento, americano, embaixada, assange, fbi, redar, snowden, eua, comunicação, material, federal
21	**Arab countries**	país, egito, egípcio, turco, turquia, saúde, árabe, mubarak, presidente, muçulmano, arábio, cairo, irmandade, erdogan, político, contra, ditador, islâmico, mursi, mohamed, hosni, regime, iêmen, novo, golpe, tunísia, líder, saleh, transição, povo

22	**Colombia - FARCs - Ecuador - Peru**	colômbio, colombiano, farc, santo, presidente, equador, refém, uribe, guerrilhar, juan, paz, armada, correa, guerrilheiro, revolucionário, libertação, peru, força, país, equatoriano, bogotá, manuel, departamento, operação, peruano, el, o, álvaro, negociação, rafael
23	**Iran**	irã, nuclear, iraniano, sanção, país, teerã, arma, ahmadinejad, agência, internacional, contra, urânio, eua, negociação, energia, potência, aiea, segurança, relação, bomba, fim, atômica, nova, rússia, ocidental, enriquecimento, mahmoud, islâmico, combustível, instalação
24	**Lybia**	líbio, rebelde, cidade, gaddafi, país, força, ditador, contra, regime, capital, trípoli, kadafi, otan, al, porta-voz, líder, muammar, civil, internacional, benghazi, nacional, transição, leal, oeste, conselho, fonte, aéreo, tv, mali, norte
25	**Police - Explosion**	polícia, local, morto, policial, explosão, ferido, vítima, cidade, corpo, homem, bomba, carro, veículo, ferida, fogo, fonte, hospital, agência, autoridade, prédio, tiro, segurança, incidente, capital, ônibus, notícia, agente, porta-voz, morte, noite
26	**Family**	mulher, ano, família, criança, pai, jovem, escola, homem, filho, menino, mãe, irmão, universidade, estudante, casal, casamentar, professor, gay, idade, educação, adolescente, familiar, sexo, amigo, parente, sexual, aluno, nome, menor, velho
27	**Italy**	cargo, primeiro-ministro, político, novo, italiano, itália, chefe, ano, berlusconi, premiê, líder, gabinete, presidente, próximo, renúncia, corrupção, ministro, público, mês, membro, executivo, atual, crise, escândalo, confiança, nome, silvio, decisão, mugabe, roma

28	**International System**	país, relação, mundo, político, mudança, novo, desenvolvimento, américa, importante, maior, economia, mundial, global, grande, problema, ano, política, econômica, social, questão, econômico, comércio, cooperação, latinar, médio, comercial, sistema, eua, internacional, governo
29	**Africa - Spain**	país, sul, espanhol, áfrico, espanha, região, presidente, ano, sudão, líder, independência, africanar, somália, contra, madri, união, mandela, internacional, capital, comunidade, quênia, costa, guerra, chefe, população, repúblico, milhão, gbagbo, último, conflito
30	**Petroleum - Energy**	milhão, us, bilião, dinheiro, r, empresa, petróleo, preço, produto, maior, ano, setor, produção, dólar, valor, trabalhador, alimento, indústria, negócio, recurso, energia, companhia, exportação, empresário, salário, gás, funcionário, alto, investimento, comercial
31	**Unknown 2**	novo, ano, cidade, york, negro, evento, lugar, festa, local, prefeito, lado, foto, noite, grande, loja, mundo, história, hotel, museu, músico, imagem, famoso, animal, carro, nome, pequeno, branco, cerimônia, futebol, mão
32	**Year - Month**	ano, mês, número, último, maior, alto, setembro, julho, janeiro, período, junho, abril, outubro, março, anterior, agostar, dezembro, maio, dado, novembro, início, seis, cincar, final, relatório, dez, ponto, fevereiro, divulgado, próximo
33	**EU**	europeu, britânico, união, europa, país, alemanha, alemão, ue, londres, merkel, cameron, bloco, chanceler, david, primeiro-ministro, trabalhista, líder, conservador, grãbretanha, berlim, brown, angela, relação, ano, bruxelas, saído, França, premiê, membro, decisão

34	**Venezuela - Cuba**	chávez, venezuela, presidente, venezuelano, cubar, cubano, hugo, caracas, país, oposição, nacional, líder, político, opositor, fidel, havano, nicolás, raúl, povo, assembleia, capriles, revolução, contra, dissidente, ano, oficial, vice-presidente, estatal, lópez, chavismo
35	**Israel - Palestine**	israel, palestino, israelense, gaza, hamas, israelenses, paz, faixar, netanyahu, território, negociação, abbas, jerusalém, cisjordânia, autoridade, judeu, contra, foguete, judaico, árabe, primeiro-ministro, assentamento, construção, médio, líder, mahmoud, lado, binyamin, fatah, anp
36	**Chile - France**	francês, frança, presidente, chile, sarkozy, chileno, hollande, mineiro, ano, sérvio, nicolas, le, françois, santiago, socialista, piñera, kosovo, bachelet, strauss-kahn, preso, belga, mês, momento, interior, trabalhador, hora, bélgica, palácio, operação, marinar
37	**North Korea**	norte, coreia, sul, navio, japão, kim, país, japonês, mar, príncipe, pyongyang, míssil, embarcação, agência, norte-coreano, lançamento, barco, costa, guerra, regime, seul, tensão, real, marinhar, exercício, oficial, marítimo, notícia, william, fonte
38	**Iraq**	iraque, terrorista, al, islâmico, iraquiano, contra, muçulmano, qaeda, xiita, atentado, segurança, bagdá, radical, sunita, grupo, militante, terrorismo, extremista, cidade, redar, país, norte, mesquita, curdo, violência, morto, alvo, membro, islã, província
39	**Riot - Strikes**	contra, manifestante, protesto, manifestação, rua, violência, polícia, policial, cidade, capital, mil, praça, centena, segurança, confronto, local, dezena, greve, jovem, noite, gás, oposição, grupo, principal, estudante, onda, ordem, violento, multidão, frente

40 **Army - War**	exército, força, soldado, guerra, civil, operação, tropa, defeso, arma, base, contra, general, região, fronteirar, armada, segurança, morto, conflito, aéreo, área, retirado, sul, comandante, missão, alvo, território, ofensivo, unidade, armado, oficial

A.2 Terms when $k = 50$

Table 10: Newspaper Topics ($k = 50$)

	Topic	Keys (Top 30)
1	**USA**	novo, york, americano, estado, negro, eua, times, new, federal, ano, califórnio, unido, universidade, branco, kennedy, governador, o, departamento, setembro, michael, washington, the, condado, john, angeles, canadense, prefeito, canadá, center, texas
2	**Natural Disasters**	terremoto, região, japão, japonês, tremor, vítima, grau, cidade, quilômetros, local, morto, dano, tsunami, magnitude, casa, indonésio, km, capital, forte, área, agência, número, país, autoridade, epicentro, província, tóquio, brasília, escombro, serviço
3	**Ocean - Hurricane**	região, navio, água, costa, chuva, tempestade, autoridade, furacão, km, forte, casa, sul, embarcação, mar, área, barco, vento, inundação, norte, emergência, terra, h, nacional, passagem, próximo, morto, quilômetros, marinhar, filipina, noite
4	**Cuba**	cubar, cubano, ano, político, preso, fidel, país, havano, raúl, libertação, líder, dissidente, revolução, viagem, fim, fome, mês, relação, comunista, irmão, regime, oficial, cincar, greve, condição, prisioneiro, povo, último, mudança, miami
5	**Unknown 1**	ano, the, história, mundo, animal, nome, famoso, museu, músico, imagem, foto, mão, show, artista, of, lugar, homem, bom, and, cantor, michelle, arte, tempo, livro, cultura, época, jovem, primeira-dama, lado, sala

6	**USA Election**	republicano, obama, democrata, trump, campanha, hillary, senador, mccain, clinton, eleitor, romney, barack, branco, john, estado, primário, presidencial, eleição, donald, delegado, presidência, convenção, contra, vitória, presidente, pesquisa, rival, novembro, governador, político
7	**Royal Family - UK**	ano, família, príncipe, cerimônia, evento, real, aniversário, sul, oficial, festa, homenagem, palácio, mundo, morte, mandela, william, rei, áfrico, futebol, rainho, grande, líder, expresidente, público, elizabeth, convidado, ocasião, momento, casamentar, bandeirar
8	**USA**	americano, eua, obama, estado, unido, presidente, washington, barack, branco, secretário, bush, george, americana, departamento, guerra, w, clinton, kerry, hillary, defeso, pentágono, john, base, congresso, relação, administração, norte-americano, viagem, gatar, segurança
9	**Economy - IMF**	banco, país, crise, economia, euro, financeiro, dívida, bilião, zonar, medida, grécia, central, europeu, público, econômica, grego, crescimento, fiscal, fmi, título, monetário, finanças, pacote, pib, novo, internacional, mercado, prazo, moeda, austeridade
10	**Money - Spain**	milhão, us, bilião, dinheiro, r, ano, espanhol, espanha, empresa, valor, público, dólar, maior, pagamento, negócio, salário, recurso, empresário, fundo, doação, funcionário, total, orçamentar, madri, financeiro, gasto, financiamento, fiscal, conta, rajoy
11	**Catholic Church**	igreja, católico, religioso, vaticano, bento, deus, francisco, cristão, santo, paulo, padrar, cardeal, mundo, fé, fiel, religião, homem, pontífice, ano, joão, novo, missar, pedrar, abuso, 2º, roma, amor, comunidade, grande, sexual

12	**Unknown 2**	político, guerra, tempo, mundo, grande, fato, tradução, próprio, problema, questão, história, claro, maneiro, difícil, verdade, lugar, sociedade, exemplo, importante, lado, momento, coisa, bom, democracia, razão, diferente, único, modo, pontar, nenhum
13	**Riot - Strikes**	manifestante, protesto, contra, manifestação, polícia, rua, violência, policial, praça, mil, capital, oposição, cidade, segurança, confronto, centena, estudante, opositor, dezena, grupo, jovem, gás, principal, greve, violento, distúrbio, onda, ordem, noite, multidão
14	**Mercosul - Venezuela**	chávez, venezuela, argentino, presidente, venezuelano, cristina, kirchner, hugo, caracas, país, oposição, aires, buenos, paraguai, nicolás, lugo, uruguai, nacional, opositor, mercosul, macri, capriles, paraguaio, ex-presidente, assembleia, presidência, unasul, clarín, uruguaio, o
15	**Year - Month**	ano, mês, número, maior, último, alto, setembro, dado, abril, outubro, julho, período, anterior, março, janeiro, maio, junho, relatório, agostar, novembro, dezembro, seis, menor, divulgado, ponto, fevereiro, início, nível, final, dez
16	**Russia**	rússia, russo, putin, ucrânia, presidente, região, país, ucraniano, separatista, vladimir, soviético, geórgia, medvedev, kremlin, conflito, território, kiev, sul, crimeia, líder, contra, independência, otan, união, ministério, autoridade, guerra, ocidental, lavrov, ossétia
17	**Political Leader**	presidente, político, novo, líder, cargo, ano, chefe, primeiro-ministro, gabinete, mandatar, eleito, atual, próximo, nacional, presidência, corrupção, posse, principal, renúncia, premiê, decisão, nome, política, membro, ministro, executivo, vice-presidente, crise, fim, mudança

18	**Local places**	cidade, local, morador, carro, capital, trem, hotel, bairro, ônibus, noite, turista, veículo, estação, pontar, rua, área, prefeito, loja, lado, lugar, motorista, hora, casa, grande, habitante, manhã, prédio, caminhão, pequeno, estradar
19	**France**	francês, frança, sarkozy, presidente, hollande, contra, nicolas, le, jornal, françois, ano, socialista, frente, strauss-kahn, belga, bélgica, marinar, público, novo, interior, maio, ministério, nacional, pen, próximo, mês, direito, charlie, momento, hotel
20	**Crime - Police**	polícia, policial, homem, morto, corpo, vítima, local, morte, agente, autoridade, arma, investigação, índio, tiro, indiano, assassinato, crime, encontrado, segurança, atirador, fogo, incidente, cidade, suspeito, jovem, carro, noite, criminoso, violência, terrorista
21	**News**	jornal, jornalista, tv, redar, vídeo, imagem, site, canal, internet, mensagem, mídia, foto, televisão, emissor, comunicação, notícia, informação, twitter, página, reportagem, facebook, repórter, social, rádio, publicado, público, texto, divulgado, comentário, publicação
22	**Brazilian Presidents - Latin America**	brasil, brasileiro, presidente, país, lula, dilma, chile, bolívia, morales, boliviano, embaixador, peru, relação, chileno, itamaraty, luiz, evo, o, brasília, embaixada, rousseff, peruano, paulo, inácio, região, ministério, patriota, amorim, piñera, américa
23	**Lybia**	líbio, rebelde, gaddafi, cidade, país, força, ditador, contra, trípoli, regime, kadafi, capital, civil, otan, muammar, conselho, benghazi, al, leal, líder, aéreo, internacional, porta-voz, oeste, transição, nacional, zonar, cnt, operação, arma

24	**UK**	britânico, londres, jornal, the, cameron, david, trabalhista, primeiro-ministro, news, ano, conservador, brown, premiê, grã-bretanha, bbc, inglaterra, inglês, australiano, austrália, of, escândalo, escócio, blair, guardian, murdoch, irlanda, world, decisão, gordon, libra
25	**Immigrants - Refugees**	país, imigrante, refugiado, guerra, áfrico, sul, região, milhão, fronteirar, imigração, ano, ilegal, sérvio, conflito, sudão, independência, população, internacional, maioria, somália, africanar, estrangeiro, situação, número, comunidade, fronteira, kosovo, quênia, violência, capital
26	**UN - Security Council**	onu, país, unido, conselho, nação, segurança, internacional, organização, haiti, missão, resolução, humanitário, membro, secretário-geral, haitiano, ban, situação, paz, estado, comunidade, embaixador, ki-moon, representante, órgão, geral, príncipe, civil, permanente, necessidade, diplomata
27	**Intelligence - Spy - Wikileaks**	informação, serviço, inteligência, documento, agência, segurança, funcionário, dado, secreto, agente, investigação, espionagem, site, wikileaks, americano, nacional, eua, embaixada, assange, departamento, acesso, relatório, autoridade, snowden, internet, nsa, fonte, revelação, jornal, cidadão
28	**Army - War**	exército, força, soldado, operação, civil, tropa, guerra, defeso, general, morto, armada, base, contra, região, comandante, armado, norte, conflito, retirado, segurança, área, oficial, sul, ofensivo, unidade, missão, helicóptero, fronteirar, aéreo, arma
29	**Israel - Palestine**	israel, palestino, israelense, gaza, hamas, israelenses, faixar, paz, netanyahu, território, abbas, jerusalém, cisjordânia, judeu, autoridade, contra, foguete, negociação, primeiro-ministro, judaico, assentamento, árabe, construção, médio, binyamin, islâmico, mahmoud, fatah, anp, premiê

30	**Law - Constitution**	lei, projeto, congresso, decisão, medido, câmara, senado, novo, deputado, aprovação, aprovado, texto, nacional, maioria, constituição, medida, público, legislação, comissão, favor, sistema, supremo, legal, constitucional, contra, federal, vigor, proibição, legislativo, mudança
31	**EU**	europeu, união, país, europa, alemanha, alemão, ue, bloco, merkel, chanceler, berlim, angela, bruxelas, membro, frança, comissão, suíço, holanda, nazista, ano, itália, espanha, áustria, holandês, líder, chefe, continente, maior, saído, decisão
32	**Global Health - WHO**	médico, hospital, caso, doença, país, morte, gripar, tratamento, vírus, número, suíno, novo, paciente, organização, contra, oms, confirmado, dor, mundial, méxico, autoridade, sintoma, exame, epidemiar, cincar, ministério, infecção, estado, h1n1, comum
33	**Explosion**	local, explosão, agência, ferido, fonte, morto, notícia, bomba, vítima, ferida, cidade, porta-voz, press, efe, capital, associated, presse, incêndio, france, ministério, reuters, fogo, oficial, brasília, autoridade, hospital, prédio, cincar, número, seis
34	**Judicial System**	prisão, tribunal, acusação, crime, contra, ano, juiz, morte, julgamento, acusado, investigação, decisão, promotor, detenção, advogado, preso, assassinato, condenado, sentença, judicial, audiência, defeso, ordem, promotoria, autoridade, condenação, prisioneiro, caso, penal, prova
35	**Election**	eleição, eleitoral, voto, votação, oposição, presidencial, vitória, campanha, partido, político, maioria, eleitor, coalizão, resultado, candidato, urna, turno, pesquisa, nacional, cadeira, pleito, conservador, líder, esquerdar, legislativo, popular, frente, deputado, reeleição, democrático

36	**Iraq**	iraque, terrorista, al, islâmico, iraquiano, qaeda, contra, xiita, atentado, bagdá, segurança, bin, sunita, redar, terrorismo, laden, radical, extremista, militante, grupo, país, alqaeda, líder, alvo, osama, norte, membro, americano, milícia, operação
37	**Honduras**	presidente, país, zelaya, golpe, honduras, crise, oea, interino, político, embaixada, internacional, junho, micheletti, hondurenho, organização, novembro, nacional, estado, manuel, congresso, presidência, costa, eleição, rico, próximo, constituição, josé, supremo, pressão, roberto
38	**Aviation**	avião, aeroporto, aéreo, passageiro, companhia, voo, aeronave, autoridade, brasília, segurança, porta-voz, local, hora, internacional, aviação, airlines, informação, problema, agência, horário, pistar, helicóptero, carga, tráfego, air, incidente, operação, nenhum, investigação, piloto
39	**Colombia - FARCs - Ecuador - Drug**	colômbio, colombiano, farc, méxico, santo, equador, mexicano, droga, uribe, guerrilhar, refém, correa, guerrilheiro, revolucionário, juan, armada, presidente, el, equatoriano, bogotá, libertação, força, paz, narcotráfico, manuel, tráfico, operação, fronteirar, o, cartel
40	**Afghanistan - Pakistan - Terrorism**	afeganistão, paquistão, afegão, país, taleban, paquistanês, contra, otan, província, segurança, insurgente, tropa, força, civil, região, militante, karzai, morte, porta-voz, talebans, internacional, soldado, cabul, terrorista, islâmico, morto, distrito, tribal, fronteirar, operação
41	**Energy**	petróleo, energia, água, produto, preço, produção, trabalhador, gás, usina, mineiro, alimento, fábrica, combustível, nível, maior, central, setor, problema, terra, produtor, tonelada, nuclear, fukushima, sistema, alto, indústria, estatal, metro, vazamento, operação

42	**China - Italy**	chinar, chinês, italiano, itália, berlusconi, pequim, ano, líder, comunista, oficial, primeiro-ministro, autoridade, hong, província, país, tibetano, kong, xinhua, premiê, taiwan, silvio, relação, região, xi, hu, roma, tibete, estatal, último, vietnã
43	**Family**	mulher, ano, criança, família, pai, escola, jovem, menino, filho, mãe, homem, estudante, irmão, casal, professor, gay, casamentar, idade, universidade, sexual, educação, sexo, homossexual, adolescente, menor, aluno, mês, amigo, familiar, parente
44	**International System**	país, mundo, economia, maior, desenvolvimento, ano, mundial, comércio, mudança, américa, global, econômica, grande, novo, social, econômico, comercial, médio, problema, político, latinar, política, população, investimento, importante, crescimento, sistema, relação, pobreza, década
45	**Human Rights**	direito, humano, contra, país, liberdade, internacional, organização, político, violação, paz, prêmio, comissão, democracia, nobel, ano, civil, relatório, ativistas, mianmar, ong, grupo, ativista, expressão, entidade, situação, regime, comitê, democrático, crítica, sociedade
46	**North Korea**	norte, coreia, sul, país, míssil, kim, nuclear, pyongyang, guerra, regime, defeso, lançamento, norte-coreano, agência, tensão, unido, japão, exercício, seul, estado, contra, novo, chinar, sul-coreano, satélite, segurança, longo, comunista, eua, mar
47	**Egypt**	muçulmano, egito, país, egípcio, árabe, saúde, mubarak, islâmico, cairo, arábio, irmandade, iêmen, contra, mohamed, mursi, ditador, hosni, al, islã, regime, tunísia, presidente, saleh, islamita, mohammed, revolução, religioso, mesquita, abdullah, autoridade

48	**Syria - Turkey**	sírio, assad, regime, turquia, turco, rebelde, contra, bashar, país, damasco, arma, oposição, conflito, líbano, civil, cidade, árabe, erdogan, libanês, guerra, ditador, grupo, força, al, química, morto, curdo, região, violência, fronteirar
49	**Iran**	irã, nuclear, iraniano, sanção, país, teerã, arma, ahmadinejad, internacional, agência, urânio, contra, eua, potência, segurança, aiea, energia, conselho, atômica, unido, bomba, enriquecimento, islâmico, mahmoud, onu, nova, fim, ocidental, estado, negociação
50	**International negotiation**	relação, reunião, negociação, país, diálogo, exterior, declaração, cúpula, paz, assunto, representante, chanceler, questão, solução, líder, fim, conversa, cooperação, ambos, bilateral, diplomático, conflito, posição, compromisso, presidente, parte, decisão, conferência, discussão, lado

A.3 Terms when $k = 80$ – Chosen model

Table 11: Newspaper Topics ($k = 80$)

	Topic	Keys (Top 30)
1	**UN - Security Council**	onu, conselho, segurança, unido, nação, internacional, organização, resolução, membro, país, estado, secretário-geral, ban, reunião, contra, comunidade, missão, geral, permanente, ki-moon, sanção, novo, órgão, representante, texto, declaração, embaixador, ação, diplomata, assembleia
2	**Lybia**	líbio, gaddafi, rebelde, ditador, cidade, trípoli, regime, força, país, kadafi, otan, contra, capital, muammar, benghazi, al, civil, conselho, aéreo, leal, líder, transição, cnt, internacional, oeste, nacional, porta-voz, misrata, muamar, tv
3	**Finance**	us, milhão, bilião, dinheiro, r, ano, valor, dólar, público, empresa, pagamento, fundo, orçamentar, doação, recurso, financeiro, total, maior, gasto, financiamento, negócio, salário, fiscal, conta, empresário, verba, rico, banco, funcionário, doador
4	**Immigration (Australia, Thailand, Indonesia)**	país, imigrante, ilegal, imigração, ano, estrangeiro, austrália, tailândia, indonésio, cidadão, australiano, origem, maioria, novo, tailandês, mês, camisa, cidadania, autoridade, bancoc, entrado, deportação, cincar, número, migratório, residência, principal, maior, milhão, último
5	**USA cities**	novo, york, negro, americano, arma, estado, eua, federal, califórnio, kennedy, governador, boston, condado, prefeito, unido, michael, branco, angeles, new, john, departamento, times, cidade, o, texas, jersey, estadual, center, martin, bloomberg

6	**Natural Disasters**	região, terremoto, cidade, casa, tremor, chuva, forte, grau, dano, área, quilômetros, km, província, magnitude, autoridade, água, sul, inundação, emergência, morador, serviço, tsunami, local, morto, vítima, tempestade, atingido, onda, terra, epicentro
7	**Uruguay - Malvinas**	país, região, território, sul, conflito, ano, guerra, população, soberania, frente, ilha, último, uruguai, tensão, regional, área, lado, uruguaio, presença, mujica, maioria, independência, amplo, fim, vizinho, década, habitante, malvinas, ambos, territorial
8	**Global Health - WHO**	médico, doença, caso, hospital, gripar, vírus, país, morte, tratamento, suíno, paciente, novo, oms, número, organização, sintoma, dor, contra, mundial, confirmado, infecção, h1n1, exame, epidemiar, estado, ebola, respiratório, autoridade, infectado, comum
9	**Mexico**	méxico, mexicano, droga, tráfico, cidade, el, fronteirar, cartel, o, criminoso, autoridade, crime, violência, san, narcotráfico, estado, traficante, federal, calderón, cocaína, polícia, ano, unido, país, salvador, felipe, peña, segurança, município, guatemala
10	**Syria - Lebanon**	sírio, assad, regime, bashar, damasco, rebelde, arma, oposição, árabe, líbano, país, conflito, civil, libanês, contra, ditador, guerra, al, cidade, grupo, química, força, violência, opositor, observatório, homs, onu, aleppo, repressão, síria
11	**Death - Victims - Survivors**	morto, vítima, corpo, morte, número, local, hospital, autoridade, sobrevivente, tragédia, encontrado, ferido, oficial, familiar, família, momento, cidade, resto, noite, identificado, parente, funeral, oito, levado, cadáver, mortal, informação, médico, dezena, nove

12	**International negotiation**	negociação, reunião, diálogo, paz, fim, solução, conversa, parte, líder, lado, representante, declaração, conflito, final, conferência, cúpula, questão, compromisso, ambos, discussão, próximo, possível, delegação, presidente, condição, negociador, mesa, esforço, objetivo, prazo
13	**Judicial System**	tribunal, prisão, juiz, acusação, contra, julgamento, decisão, ano, crime, judicial, supremo, acusado, advogado, sentença, audiência, promotor, defeso, ordem, promotoria, corrupção, condenação, judiciário, recurso, caso, condenado, legal, processo, perpétuo, penal, extradição
14	**USA primary election**	obama, republicano, democrata, campanha, senador, mccain, romney, eleitor, hillary, barack, john, estado, primário, delegado, branco, convenção, presidencial, eleição, clinton, contra, presidência, bush, governador, ponto, flórido, pesquisa, rival, novembro, vitória, candidato
15	**Honduras**	presidente, zelaya, país, honduras, golpe, interino, crise, oea, político, micheletti, hondurenho, internacional, manuel, junho, organização, eleição, congresso, embaixada, presidência, novembro, rico, nacional, roberto, estado, josé, nicarágua, constituição, supremo, tegucigalpa, lobo
16	**Germany - Spain**	alemão, espanhol, alemanha, espanha, merkel, chanceler, berlim, ano, madri, angela, nazista, país, rajoy, eta, europa, basco, p, catalunha, judeu, barcelona, zapatero, mariano, regional, catalão, hitler, independência, holocausto, chefe, concentração, oriental
17	**Chile**	estudante, universidade, chile, escola, professor, chileno, mineiro, ano, educação, santiago, aluno, piñera, trabalhador, bachelet, aula, universitário, san, preso, josé, metro, michelle, mês, hora, local, sebastián, ciência, faculdade, instituição, pinochet, público

18	**Crime - Murders**	morte, assassinato, crime, ano, homem, contra, vítima, execução, polícia, mulher, acusação, condenado, strauss-kahn, agressão, homicídio, tentativo, sexual, suicídio, sakineh, criminoso, quartar, acusado, violência, jovem, executado, hotel, mão, apedrejamento, prisão, noite
19	**Russia**	russo, rússia, putin, ucrânia, presidente, ucraniano, vladimir, soviético, separatista, geórgia, medvedev, região, kremlin, kiev, crimeia, otan, líder, ossétia, dmitri, país, lavrov, união, yanukovich, ocidental, conflito, polônia, território, georgiano, ocidente, polonês
20	**USA conflicts**	eua, guerra, americano, bush, defeso, george, iraque, base, w, tropa, afeganistão, ano, pentágono, segurança, americana, soldado, estratégia, secretário, retirado, país, general, setembro, washington, força, gatar, conflito, unido, aliado, estado, presidente
21	**Egypt**	egito, egípcio, mubarak, país, cairo, irmandade, árabe, muçulmano, presidente, mursi, hosni, ditador, tunísia, mohamed, regime, novo, exército, protesto, força, revolução, político, transição, islamita, ben, povo, praça, tahrir, morsi, popular, islâmico
22	**Venezuela**	chávez, venezuela, venezuelano, presidente, hugo, caracas, oposição, nacional, nicolás, opositor, assembleia, país, líder, capriles, povo, contra, chavismo, bolívar, câncer, lópez, chavista, democrático, vice-presidente, mud, henrique, unidade, golpe, bolivariana, estatal, chavistas
23	**International Press**	agência, notícia, fonte, efe, reuters, press, porta-voz, associated, france, presse, oficial, celular, informação, wap.folha.com.br, próximo, citado, local, madeleine, ministério, mccann, português, maio, polícia, momento, entanto, ap, anonimato, desaparecimento, funcionário, último

24	**Royal Family - UK**	ano, príncipe, cerimônia, real, evento, aniversário, família, festa, rei, william, palácio, rainho, homenagem, oficial, elizabeth, convidado, charlar, casamentar, público, kate, celebração, primeira-dama, comemoração, britânico, casal, michelle, ocasião, diana, 2ª, viagem
25	**International System**	país, mundo, américa, desenvolvimento, relação, mundial, global, latinar, internacional, novo, cooperação, comércio, mudança, eua, comercial, maior, economia, importante, econômica, econômico, grande, cúpula, governo, política, nação, acordo, investimento, político, região, área
26	**Iran**	irã, nuclear, iraniano, teerã, sanção, país, ahmadinejad, arma, urânio, internacional, eua, potência, agência, contra, aiea, energia, enriquecimento, atômica, mahmoud, islâmico, bomba, ocidental, presidente, rússia, repúblico, nova, pacífico, ocidente, fim, regime
27	**Argentina**	argentino, cristina, kirchner, presidente, aires, buenos, macri, província, ex-presidente, clarín, político, néstor, o, ditadura, fernández, último, público, nisman, principal, mauricio, jorge, scioli, daniel, frente, federal, rosado, carlos, oficial, promotor, presidência
28	**Terrorism**	polícia, policial, explosão, bomba, local, morto, ferido, homem, segurança, cidade, carro, explosivo, ferida, veículo, tiro, prédio, incidente, terrorista, agente, capital, vítima, atentado, interior, hospital, fogo, atirador, fonte, autoridade, ação, cincar
29	**Family links**	ano, família, criança, pai, jovem, menino, filho, mãe, irmão, mulher, escola, idade, adolescente, menor, amigo, velho, parente, familiar, homem, avô, mês, adulto, casal, neto, cincar, nome, tempo, seis, idoso, infância

30	**Haiti**	haiti, país, missão, terremoto, haitiano, onu, capital, príncipe, brasileiro, organização, reconstrução, internacional, nação, unido, civil, janeiro, paz, número, morto, minustah, humanitário, dominicano, grande, república, vítima, cólera, situação, médico, mil, tragédia
31	**Unknown 1**	local, brasília, incêndio, cidade, trem, horário, hora, ônibus, fogo, bombeiro, estação, noite, veículo, autoridade, manhã, capital, região, motorista, pontar, chama, caminhão, linha, serviço, metrô, ferido, próximo, fechado, km, emergência, porta-voz
32	**Investigation**	investigação, informação, serviço, relatório, inteligência, agente, funcionário, autoridade, segurança, suspeito, operação, secreto, polícia, departamento, fbi, comissão, investigador, diretor, evidência, prova, envolvimento, acusação, suspeita, agência, nome, interrogatório, federal, ligação, membro, alto
33	**Human Rights**	direito, humano, contra, liberdade, internacional, organização, país, civil, violação, comissão, político, relatório, lei, grupo, expressão, ong, entidade, ativistas, anistia, situação, defeso, cidadão, abuso, democracia, violência, sociedade, crítica, repressão, defensor, independente
34	**Gender - Gay**	mulher, homem, lei, ano, casamentar, gay, direito, sexual, sexo, homossexual, casal, público, contra, social, sociedade, feminino, igualdade, gênero, discriminação, véu, proibição, tipo, prático, questão, civil, união, conservador, identidade, orientação, casamento
35	**Social Media - Famous People**	vídeo, imagem, foto, mensagem, redar, internet, site, twitter, facebook, social, página, mundo, tv, campanha, músico, nome, artista, show, gravação, comentário, minuto, cantor, famoso, youtube, fotografia, público, palavra, ator, mídia, história

36	**Diplomacy**	relação, exterior, país, ministério, embaixada, embaixador, diplomata, diplomático, assunto, funcionário, porta-voz, declaração, diplomática, chanceler, situação, oficial, reunião, representante, representação, viagem, canadá, diplomacia, consulado, autoridade, bilateral, canadense, respostar, decisão, internacional, laço
37	**Africa - South Africa**	país, áfrico, ano, sul, paz, presidente, prêmio, nobel, líder, mundo, mandela, nacional, ex-presidente, africanar, contra, negro, africano, último, fim, branco, zuma, sul-africano, nelson, continente, político, democracia, comitê, povo, apartheid, carter
38	**Soccer**	ano, animal, maconha, futebol, clube, mundo, jogo, local, breivik, norueguês, carnar, noruega, restaurante, estádio, público, maior, oslo, jogador, álcool, grande, problema, tipo, cachorro, parque, lugar, time, jovem, dono, copar, cão
39	**Law - Constitution**	lei, congresso, projeto, câmara, senado, deputado, presidente, novo, aprovação, maioria, aprovado, texto, constituição, oposição, legislativo, medido, representante, favor, legislação, votação, constitucional, senador, comissão, sessão, legislador, nacional, federal, sistema, republicano, executivo
40	**Intelligence - Spy - Wikileaks**	informação, documento, dado, site, wikileaks, americano, eua, segurança, espionagem, internet, assange, agência, serviço, acesso, nacional, snowden, comunicação, inteligência, nsa, secreto, revelação, sistema, vigilância, estado, vazamento, fundador, suécio, computador, confidencial, tecnologia
41	**China**	chinar, chinês, pequim, comunista, província, autoridade, oficial, ano, hong, líder, tibetano, xinhua, kong, taiwan, xi, hu, tibete, estatal, liu, dalai-lama, asiático, chen, vietnã, maior, central, wen, jinping, região, bo, jintao

42	**Turkey**	turquia, turco, país, contra, erdogan, curdo, presidente, primeiro-ministro, istambul, gbagbo, costa, ancara, pkk, fronteirar, tayyip, recep, tentativo, golpe, marfim, ouattara, território, ação, último, chefe, lado, premiê, maior, autoridade, principal, trabalhador
43	**USA elections**	trump, hillary, clinton, campanha, eua, presidente, donald, americano, secretário, branco, republicano, político, bill, democrata, sanders, washington, assessor, cruz, declaração, presidência, ex-presidente, grande, bilionário, presidencial, york, empresário, eleito, magnata, posição, new
44	**Conflicts - Genocide**	guerra, país, internacional, sérvio, sudão, sul, independência, contra, kosovo, crime, quênia, conflito, região, genocídio, capital, repúblico, ano, sudanês, presidente, congo, norte, bósnio, paz, tpi, violência, humanidade, africanar, onu, darfur, civil
45	**India**	índio, hotel, indiano, novo, turista, local, capital, país, cidade, região, ano, cincar, último, estrangeiro, terrorista, turismo, mumbai, autoridade, turístico, principal, ponto, próximo, segurança, dez, seis, oficial, grande, atentado, distrito, agência
46	**Unknown 2**	decisão, medida, medido, nova, novo, anúncio, fim, sistema, nacional, restrição, ações, contra, lista, regra, tomado, objetivo, proibição, entanto, ação, país, suspensão, necessário, respostar, norma, vigor, pressão, público, possibilidade, apesar, ordem
47	**Israel - Palestine**	israel, palestino, israelense, gaza, hamas, israelenses, faixar, netanyahu, abbas, território, paz, jerusalém, cisjordânia, judeu, autoridade, árabe, assentamento, primeiro-ministro, foguete, judaico, médio, construção, binyamin, fatah, mahmoud, anp, islâmico, contra, premiê, palestina

48	**Petroleum**	petróleo, preço, produto, país, empresa, produção, energia, setor, gás, maior, indústria, exportação, fábrica, alimento, estatal, trabalhador, companhia, comércio, comercial, produtor, importação, dólar, combustível, petrolífero, reserva, problema, terra, grande, gasolina, principal
49	**Political Leaders**	político, líder, eleição, partido, coalizão, novo, conservador, esquerdar, maioria, política, liberal, direito, oposição, nacional, cadeira, democracia, popular, aliançar, democrático, ano, formação, principal, atual, primeiro-ministro, premiê, analista, socialista, social, reforma, liderança
50	**North Korea**	norte, coreia, sul, míssil, país, kim, nuclear, pyongyang, regime, norte-coreano, lançamento, seul, mianmar, exercício, guerra, sul-coreano, agência, tensão, japão, comunista, satélite, defeso, unido, sul-coreana, chinar, península, kyi, lee, suu, foguete
51	**Catholic Church**	igreja, católico, vaticano, bento, francisco, deus, santo, religioso, padrar, cardeal, paulo, pontífice, cristão, joão, fé, fiel, missar, mundo, abuso, 2º, pedrar, bispar, bispo, sexual, arcebispo, sacerdote, novo, cristo, amor, jesus
52	**News**	jornal, jornalista, tv, canal, times, emissor, mídia, reportagem, comunicação, the, repórter, televisão, informação, redar, publicado, new, rádio, post, diário, york, publicação, edição, notícia, site, artigo, cobertura, cnn, público, texto, veículo
53	**Afghanistan - Pakistan**	afeganistão, paquistão, afegão, taleban, paquistanês, país, otan, província, insurgente, contra, força, tropa, karzai, segurança, talebans, civil, cabul, militante, soldado, internacional, porta-voz, tribal, distrito, islamabad, musharraf, região, morte, operação, hamid, islâmico

54	**Islam - Terrorism**	islâmico, muçulmano, contra, terrorista, religioso, radical, mesquita, grupo, extremista, cristão, islã, terrorismo, religião, atentado, nigéria, comunidade, violência, militante, mundo, maioria, haram, boko, maomé, nigeriano, país, alvo, líder, profeta, maior, nome
55	**Elections**	eleição, eleitoral, voto, votação, presidencial, campanha, eleitor, vitória, candidato, resultado, urna, turno, pesquisa, pleito, oposição, milhão, reeleição, maioria, participação, presidente, colégio, comissão, ponto, legislativo, nacional, número, candidatura, mandatar, vantagem, opositor
56	**Peru - Bolivia**	bolívia, morales, presidente, boliviano, peru, o, evo, país, peruano, departamento, indígena, paz, região, cruz, governador, santo, nacional, garcía, fujimori, humala, contra, regional, opositor, ex-presidente, político, camponês, novo, alberto, oposição, autonomia
57	**Iraq**	iraque, iraquiano, xiita, bagdá, sunita, al, cidade, segurança, país, força, norte, contra, província, violência, curdo, hussein, saddam, região, islâmico, capital, maliki, tropa, milícia, exército, invasão, maioria, americano, mossul, coalizão, sectário
58	**Italy**	italiano, itália, berlusconi, primeiro-ministro, premiê, ano, roma, silvio, chefe, milão, monti, jovem, menor, jornal, napolitano, próximo, festa, noite, mês, escândalo, o, máfia, político, contra, mario, liberdade, momento, último, líder, italiana
59	**Mercosul - Unasul**	brasil, brasileiro, presidente, lula, dilma, país, paraguai, mercosul, lugo, itamaraty, luiz, chanceler, rousseff, paraguaio, inácio, paulo, brasília, fernando, amorim, relação, bloco, patriota, reunião, senador, embaixador, unasul, presidência, posição, josé, franco

60	**Unknown 3**	cidade, morador, lado, bairro, local, carro, lugar, pequeno, rua, loja, metro, prédio, museu, construção, casa, mão, edifício, noite, ano, antigo, pé, grande, tempo, gente, ninguém, apartamento, medo, sala, roupa, frente
61	**Cuba**	cubar, cubano, ano, fidel, político, havano, raúl, dissidente, regime, líder, comunista, revolução, país, viagem, irmão, relação, oficial, fim, mudança, miami, fome, greve, opositor, abertura, sánchez, histórico, reforma, último, povo, cincar
62	**Army - Troops**	exército, força, soldado, civil, operação, contra, cidade, tropa, rebelde, morto, região, ofensivo, fronteirar, armada, arma, aéreo, área, norte, sul, armado, general, combate, defeso, confronto, alvo, ferido, militante, fonte, porta-voz, comandante
63	**Colombia - FARCs - Ecuador**	colômbio, colombiano, farc, santo, equador, uribe, presidente, guerrilhar, armada, correa, força, guerrilheiro, revolucionário, juan, bogotá, equatoriano, refém, manuel, operação, libertação, paz, álvaro, rafael, betancourt, exército, departamento, território, defeso, ingrid, acampamento
64	**EU**	europeu, união, país, europa, ue, bloco, comissão, alemanha, bruxelas, portugal, saído, suíço, França, crise, holanda, membro, português, continente, cúpula, européia, holandês, ministro, hungria, comum, lisboa, itália, decisão, reunião, irlanda, único
65	**Japan**	japão, japonês, nuclear, usina, água, nível, agência, central, energia, fukushima, tóquio, março, reatores, radiação, vazamento, tsunami, reator, terremoto, autoridade, local, sistema, país, maior, área, crise, tepco, segurança, kyodo, unidade, desastre

66	**France - Belgium**	francês, frança, sarkozy, presidente, hollande, nicolas, le, françois, socialista, belga, bélgica, frente, marinar, contra, pen, nacional, charlie, interior, bruxelas, bernard, direito, hebdo, palácio, jornal, chefe, esquerdar, público, eliseu, república, maio
67	**Unknown 4**	político, problema, questão, claro, tempo, fato, grande, importante, difícil, momento, situação, possível, especialista, posição, maneiro, opinião, mudança, próprio, tipo, diferente, nenhum, bom, exemplo, coisa, analista, pontar, algum, razão, universidade, maior
68	**Aviation**	avião, aeroporto, aéreo, companhia, passageiro, voo, aeronave, autoridade, airlines, aviação, internacional, segurança, porta-voz, pistar, air, problema, piloto, cancelado, hora, boeing, tráfego, informação, terminal, cancelamento, nuvem, cinza, helicóptero, área, tripulante, viagem
69	**EU economy**	banco, euro, financeiro, dívida, crise, país, zonar, grécia, bilião, europeu, grego, economia, central, medida, fmi, título, monetário, público, finanças, internacional, pacote, novo, mercado, austeridade, fiscal, juro, instituição, prazo, bancário, empréstimo
70	**Unknown 5**	mundo, história, político, tradução, tempo, grande, próprio, sociedade, guerra, verdade, fato, democracia, século, povo, lugar, maneiro, exemplo, década, social, realidade, único, bom, modo, palavra, lado, cultura, razão, coisa, homem, liberdade
71	**UK**	britânico, londres, the, cameron, david, of, news, brown, trabalhista, grã-bretanha, bbc, jornal, primeiro-ministro, inglaterra, premiê, conservador, escócio, blair, escândalo, murdoch, world, inglês, and, gordon, escocês, guardian, libra, irlanda, tabloide, may

72	**Prison - Guantanamo**	prisão, preso, libertação, autoridade, ano, prisioneiro, detenção, refém, cincar, mês, guantánamo, libertado, base, condição, seis, detentos, homem, sete, sequestrado, último, segurança, hora, jornalista, porta-voz, fuga, sequestrador, nenhum, suspeito, transferência, passado
73	**Economy**	ano, economia, crescimento, maior, alto, número, econômica, médio, país, população, nível, dado, econômico, social, emprego, índice, pib, menor, crise, setor, mês, inflação, redução, público, anterior, período, trimestre, economista, relatório, preço
74	**Terrorism - Al Qaeda**	al, terrorista, qaeda, saúde, bin, iêmen, laden, redar, arábio, contra, país, árabe, líder, islâmico, al-qaeda, osama, saleh, terrorismo, operação, segurança, mali, militante, abdullah, somália, organização, membro, abu, atentado, autoridade, emirado
75	**Riot - Strikes**	manifestante, protesto, contra, manifestação, rua, polícia, violência, policial, cidade, mil, praça, capital, oposição, segurança, confronto, centena, gás, jovem, greve, dezena, distúrbio, opositor, violento, força, multidão, onda, grupo, repressão, lacrimogêneo, principal
76	**Ocean - Ships**	navio, costa, mar, furacão, embarcação, barco, km, água, marinhar, tempestade, h, marítimo, pirata, vento, sul, operação, capitão, categoria, passagem, quilômetros, autoridade, naufrágio, próximo, costeiro, tropical, norte, região, direção, nacional, tripulante
77	**Refugees**	refugiado, país, humanitário, milhão, situação, número, população, fronteirar, organização, onu, internacional, região, crise, mil, nação, assistência, maior, acesso, fronteira, alimento, alto, conflito, migrante, fome, agência, condição, campo, área, último, maioria

78	**USA presidents**	americano, obama, eua, estado, unido, presidente, barack, washington, branco, departamento, secretário, kerry, americana, john, norte-americano, biden, viagem, congresso, norte-americanos, administração, povo, porta-voz, norte-americana, joe, vice-presidente, carney, esforço, assessor, gibbs, robert
79	**Year - Month**	ano, mês, janeiro, julho, junho, setembro, dezembro, outubro, maio, agostar, março, abril, novembro, fevereiro, início, último, seis, próximo, final, novo, fim, período, cincar, 1º, dez, anterior, realizado, terceirar, meado, sete
80	**President - Prime-minister**	presidente, cargo, novo, político, chefe, eleito, mandatar, gabinete, posse, líder, presidência, atual, renúncia, próximo, presidencial, nome, nacional, membro, vice-presidente, ministro, ano, crise, executivo, mudança, primeiro-ministro, transição, principal, assessor, mugabe, corrupção

B Appendix: Congress

Table 12: Congress' CFR: Database summary

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
year	62,410	Categorical	–	–	–	–	–
congress	62,410	Categorical	–	–	–	–	–
author	62,410	Categorical	–	–	–	–	–
commt_id	62,410	Categorical	–	–	–	–	–
meeting_url	62,410	Categorical	–	–	–	–	–
meeting_date	62,410	Categorical	–	–	–	–	–
meeting_name	62,410	Categorical	–	–	–	–	–
speech_raw	62,410	Categorical	–	–	–	–	–
speech	62,410	Categorical	–	–	–	–	–
meeting_observation	62,410	Categorical	–	–	–	–	–
meeting_invited	62,410	Categorical	–	–	–	–	–
meeting_code	62,410	Categorical	–	–	–	–	–
meeting_summary	62,410	Categorical	–	–	–	–	–
meeting_duration	62,410	Categorical	–	–	–	–	–
meeting_tape_duration	62,410	Categorical	–	–	–	–	–
meeting_rooms	62,410	Categorical	–	–	–	–	–
meeting_location	62,410	Categorical	–	–	–	–	–
meeting_pages	18,238	61.1	26.3	1.0	44.0	76.0	129.0
meeting_start_time	62,410	Categorical	–	–	–	–	–
meeting_end_time	62,410	Categorical	–	–	–	–	–
meeting_join	62,410	0.1	0.3	0	0	0	1
vice_president	62,410	0.000	0.02	0	0	0	1
meeting_president	62,410	0.4	0.5	0	0	1	1
ambassador	62,410	0.001	0.03	0	0	0	1
rapporteur	62,410	0.002	0.04	0	0	0	1
translator	62,410	0.000	0.02	0	0	0	1
secretary	62,410	0.000	0.01	0	0	0	1
minister	62,410	0.02	0.1	0	0	0	1
deputy	62,410	0.2	0.4	0	0	0	1
senator	62,410	0.1	0.3	0	0	0	1
governor	62,410	0.000	0.01	0	0	0	1
vice_governor	62,410	0.000	0.004	0	0	0	1
general	62,410	0.000	0.02	0	0	0	1
colonel	62,410	0.001	0.03	0	0	0	1
coordinator	62,410	0.003	0.1	0	0	0	1
military	62,410	0.003	0.1	0	0	0	1
bishop	62,410	0.000	0.01	0	0	0	1
police_chief	62,410	0.001	0.02	0	0	0	1
jugde	62,410	0.000	0.01	0	0	0	1
coalition	62,410	Categorical	–	–	–	–	–
length_wd	62,390	198.0	464.9	1.0	8.0	155.0	11,065.0
length_char	62,394	1,237.5	2,908.5	1.0	48.0	985.8	70,086.0
congress_speech_id	62,410	31,205.5	18,016.4	1	15,603.2	46,807.8	62,410
name_congress	62,410	Categorical	–	–	–	–	–
party	62,410	Categorical	–	–	–	–	–
state	62,410	Categorical	–	–	–	–	–

B.1 Terms when $k = 20$

Table 13: Politician Speech Topics ($k = 20$)

	Topic	Keys (Top 30)
1	**International System**	país, brasil, internacional, político, desenvolvimento, sociedade, mundo, nação, brasileiro, novo, importante, conferência, social, organização, mudança, unido, grande, global, papel, ano, tema, participação, política, mundial, diplomacia, sustentável, importância, conselho, nacional, fundamental
2	**Defense - Haiti Submarine**	defeso, força, armada, exército, nacional, marinhar, comandante, ministério, nuclear, orçamentar, recurso, ano, general, estratégico, projetos, brasileiro, submarino, haiti, necessidade, projeto, operação, navio, área, missão, guerra, capacidade, construção, aeronáutico, brasil, milhão
3	**Unknown 1**	questão, problema, gente, momento, pontar, importante, maneiro, claro, fato, exemplo, grande, tempo, nenhum, coisa, bom, inclusive, situação, mundo, algum, verdade, tipo, próprio, exatamente, relação, evidentemente, possível, dúvida, preocupação, lugar, discussão
4	**Crime - Police**	segurança, informação, público, federal, polícia, inteligência, crime, sistema, dado, órgão, grande, brasileiro, tráfico, policial, droga, civil, estado, internet, agência, ministério, evento, operação, serviço, atividade, redar, sociedade, janeiro, nacional, terrorismo, comunicação
5	**Trade**	brasil, estado, país, unido, comércio, chinar, negociação, união, comercial, relação, europeu, acordo, brasileiro, grande, importante, área, mercosul, produto, político, maior, omc, ano, americano, alca, européia, economia, chinês, agricultura, internacional, agrícola
6	**Committee on Foreign Relations**	comissão, relação, exterior, nacional, defeso, reunião, presidente, audiência, público, senado, câmara, ministério, assunto, membro, presidência, presença, trabalho, realização, congresso, aberto, sessão, realizado, requerimento, representante, convite, ordinário, secretariar, anterior, subcomissão, extraordinário
7	**Meeting Protocol 1**	presidente, reunião, embaixador, hora, próximo, celso, amorim, josé, minuto, assunto, inclusive, joão, encerrado, costa, sugestão, sérgio, presença, favor, palavra, marcelo, sessão, jefferson, nome, exposição, vieira, crivella, oportunidade, voto, plenário, s[a]
8	**Social Issues**	brasileiro, brasil, país, ano, situação, educação, trabalhador, família, social, universidade, mulher, número, comunidade, estado, direito, japão, médico, criança, ministério, problema, exterior, grande, escola, oportunidade, tempo, cidadão, maior, condição, imigrante, haitiano
9	**Technology - Embraer**	tecnologia, aéreo, projeto, avião, brasileiro, brasil, área, empresa, indústria, sistema, tecnológico, desenvolvimento, aeronave, ciência, nacional, defeso, satélite, base, ano, equipamento, importante, aeronáutico, embraer, lançamento, capacidade, espacial, civil, transferência, aviação, recurso
10	**Senate Committee**	senador, palavra, eduardo, suplicy, pedrar, cristovam, ordem, ana, amélia, luiz, buarque, henrique, jorge, viana, aloysio, ferraço, simon, ricardo, roberto, nunes, seguido, microfone, exa, ferreiro, azeredo, indagação, requião, josé, vanessa, tuma
11	**Democracy**	presidente, direito, povo, político, humano, contra, democracia, momento, cubar, brasileiro, lula, nenhum, democrático, história, posição, liberdade, homem, fato, manifestação, verdade, opinião, mim, cubano, inclusive, líder, episódio, colega, razão, tempo, época
12	**Latin America**	brasil, país, presidente, brasileiro, mercosul, venezuela, relação, américa, argentino, integração, sul, paraguai, bolívia, político, importante, colômbio, latinar, questão, itamaraty, região, chile, situação, uruguai, equador, momento, posição, peru, boliviano, inclusive, unasul
13	**Amazon**	região, área, amazônia, fronteirar, grande, terra, questão, problema, indígena, sul, índio, ano, água, norte, bom, cidade, território, roraima, município, população, santo, ministério, grosso, maior, brasil, comunidade, guiana, recurso, pequeno, projeto
14	**Africa**	país, brasil, cooperação, relação, ano, grande, área, brasileiro, áfrico, importante, presidente, us, bilateral, milhão, político, cultural, maior, comércio, embaixada, população, português, comercial, sul, relacionamento, língua, comunidade, presença, intercâmbio, principal, inclusive
15	**Legal Issues**	lei, internacional, nacional, projeto, constituição, direito, legislativo, convenção, jurídico, texto, congresso, legislação, cooperação, executivo, aprovação, matéria, análise, técnico, presidente, constitucional, artigo, câmara, decisão, parte, relação, legal, norma, âmbito, protocolo, tratado
16	**Meeting Protocol 2**	senador, requerimento, item, senado, relatório, federal, relator, discussão, votação, matéria, autoria, repúblico, projeto, comissão, regimentar, leitura, apreciação, relatoria, terminativo, plenário, aprovação, observação, mensagem, legislativo, ad, deliberação, incisar, hoc, constituição, dezembro
17	**Diplomacy**	embaixador, repúblico, relação, brasil, exterior, ministério, embaixada, presidente, cargo, diplomata, carreira, missão, chefe, indicação, classe, diplomático, secretário, permanente, itamaraty, ordem, conselheiro, federal, função, nome, relatório, especial, branco, instituto, maria, josé
18	**Investment - Petrobras**	ano, brasil, bilião, país, empresa, energia, investimento, milhão, produto, grande, dólar, economia, indústria, banco, produção, exportação, preço, setor, petróleo, brasileiro, valor, maior, petrobras, crescimento, gás, crise, mundo, financeiro, pib, capital
19	**International Politics**	unido, estado, país, guerra, político, paz, rússia, nação, internacional, mundo, americano, conflito, onu, iraque, conselho, israel, ano, irã, sírio, contra, novo, segurança, palestino, europeu, árabe, grande, frança, médio, relação, presidente
20	**House of Representantives Committee**	deputado, palavra, dr, carlos, presença, mesa, pergunta, exposição, prof, paulo, sr, audiência, s.exa, minuto, convidado, obrigado, representante, conosco, microfone, professor, tempo, josé, palma, autor, colega, fernando, luiz, diretor, antonio, seminário

B.2 Terms when $k = 40$

Table 14: Politician Speech Topics ($k = 40$)

	Topic	Keys (Top 30)
1	**Ambassadors**	embaixador, palavra, celso, amorim, exposição, luiz, sérgio, antonio, josé, carlos, vieira, presença, roberto, s[a], alberto, maria, itamaraty, patriota, ambos, souza, exa, guimarães, eminente, indagação, barbosa, mauro, pinheiro, machadar, figueiredo, rubens
2	**Africa - CLP**	país, áfrico, ano, presidente, brasil, português, grande, sul, população, língua, angola, milhão, relação, independência, africanar, embaixada, comunidade, costa, presença, habitante, portugal, região, africano, importante, moçambique, continente, timor, pequeno, político, povo
3	**Family - Religion**	mulher, trabalhador, família, criança, ano, santo, social, homem, brasil, morte, religioso, igreja, importante, grande, negro, atenção, paulo, tempo, católico, pai, público, bom, representante, escravo, oportunidade, mãe, nome, companheiro, condição, sociedade
4	**Committee Issues**	comissão, relação, exterior, nacional, defeso, audiência, público, ministério, câmara, senado, assunto, presidente, deputado, congresso, representante, realização, membro, convite, subcomissão, presidência, iniciativo, presença, sugestão, misto, importância, secretariar, exmo, objetivo, tema, autoridade
5	**Aviation**	informação, empresa, dado, comunicação, civil, internet, redar, sistema, varig, segurança, aviação, aeroporto, aéreo, serviço, acesso, agência, brasileiro, companhia, autoridade, setor, espionagem, anac, tipo, vôo, cidadão, público, nome, nenhum, investigação, gestão

6	**Trade**	brasil, produto, exportação, brasileiro, país, indústria, comércio, empresa, grande, us, setor, investimento, ano, economia, bilião, produção, maior, comercial, crescimento, importação, produtor, agricultura, milhão, pib, desenvolvimento, pequeno, preço, exportador, industrial, área
7	**Blocs - Multilateral Groups**	negociação, país, comércio, união, omc, alca, brasil, acordo, importante, européia, área, subsídio, comercial, serviço, agricultura, agrícola, â, regra, rodado, exemplo, grande, acesso, tema, mercosul, brasileiro, posição, negociador, europeu, maneiro, setor
8	**Venezuela - Cuba**	venezuela, político, presidente, brasileiro, cubar, democracia, democrático, posição, eleição, povo, contra, oposição, colômbio, manifestação, venezuelano, liberdade, cubano, inclusive, eleitoral, momento, situação, fato, país, episódio, regime, líder, chávez, opinião, ditadura, partido
9	**Amazon**	amazônia, região, área, fronteirar, indígena, terra, índio, problema, território, roraima, questão, norte, comunidade, população, guiana, grosso, sul, faixar, água, município, povo, cidade, amazônica, amazona, quilômetros, brasileiro, grande, manaus, acre, federal
10	**Energy - Bolivia**	energia, bolívia, brasil, petrobras, petróleo, gás, brasileiro, preço, produção, itaipu, boliviano, energético, país, paraguai, grande, questão, momento, relação, natural, milhão, condição, importante, elétrica, usina, investimento, bom, fonte, mina, contrato, maior
11	**Senator 1**	senador, palavra, pedrar, josé, relator, simon, joão, marcelo, costa, jefferson, hélio, eduardo, azeredo, crivella, péres, arthur, tuma, ribeiro, favorável, virgílio, romeu, capiberibe, sugestão, flexa, sentado, gilberto, gentileza, saturnino, mozarildo, nobre

12	**Climate Change**	desenvolvimento, país, conferência, brasil, nação, sustentável, mudança, ambiental, questão, unido, planeta, ano, brasileiro, discussão, clima, meta, objetivos, global, climático, compromisso, presidente, fórum, biodiversidade, internacional, grande, maior, mundo, emissão, mundial, evento
13	**Meeting Protocol 1**	reunião, comissão, próximo, hora, minuto, sessão, aberto, encerrado, presença, plenário, extraordinário, ordinário, trabalho, anterior, favor, número, aprovado, secreto, leitura, voto, ordem, regimental, iniciado, nome, legislativo, convite, manhã, realizado, audiência, membro
14	**Legislative x Executive Dynamic**	projeto, legislativo, texto, internacional, nacional, aprovação, matéria, cooperação, repúblico, convenção, emenda, federativo, congresso, análise, presidente, parte, câmara, técnico, relativo, acordo, artigo, relator, constituição, protocolo, tratado, informação, proposição, vigor, executivo, jurídico
15	**China - BRICs**	brasil, chinar, país, político, mundo, internacional, relação, economia, chinês, mundial, grande, unido, ano, crise, brasileiro, estado, novo, brics, crescimento, global, comercial, último, econômica, índio, econômico, europa, mudança, maior, comércio, fim
16	**Brazilian Presidents**	presidente, fernando, lula, intervenção, microfone, inclusive, colega, collor, oportunidade, ordem, henrique, tempo, pronunciamento, governador, certeza, dilma, cardoso, viagem, inaudível, gabeira, sr, mim, assunto, bom, franco, desculpa, caro, vice-presidente, sugestão, riso

17	**Crime - Police**	polícia, segurança, federal, público, inteligência, crime, policial, tráfico, grande, droga, órgão, evento, informação, operação, atividade, janeiro, terrorismo, ministério, sistema, nacional, ações, área, abin, atividades, copar, fronteirar, tipo, agência, brasileiro, sociedade
18	**Arabic Countries**	país, político, árabe, presidente, novo, líbio, egito, reunião, regime, defeso, comissão, região, brasileiro, grande, ano, senador, internacional, relação, nacional, líder, contra, tunísia, saúde, unido, brasil, islâmico, próximo, arábio, sírio, mundo
19	**Army - Defense**	defeso, força, armada, exército, nacional, comandante, general, ministério, estratégico, aeronáutico, orçamentar, brasileiro, marinhar, segurança, guerra, operação, projetos, necessidade, missão, civil, soberania, estratégia, recurso, tropa, presença, maior, situação, oficial, cibernético, fronteira
20	**Human Rights**	direito, internacional, humano, brasil, brasileiro, conselho, nação, organização, país, diplomacia, político, segurança, novo, sociedade, onu, participação, membro, compromisso, proteção, princípio, relação, civil, papel, democrático, organismo, permanente, tema, nacional, fundamental, violação
21	**North America (USA, Canada, Mexico)**	estado, unido, americano, brasil, presidente, relação, méxico, país, mundo, canadá, américa, questão, importante, trump, congresso, norte-americano, inclusive, ano, momento, brasileiro, obama, coréia, maior, mexicano, bush, exemplo, norte-americana, americana, nação, washington
22	**Unknown 1**	gente, problema, mundo, grande, bom, coisa, brasil, tempo, ninguém, ano, nenhum, exemplo, alguém, pontar, história, maneiro, verdade, época, lugar, quê, daqui, difícil, diferente, lado, dinheiro, pessoal, frente, mim, momento, claro

23	**Unknown 2**	questão, momento, fato, relação, inclusive, claro, importante, problema, evidentemente, maneiro, pontar, exatamente, preocupação, situação, nenhum, algum, dúvida, posição, próprio, verdade, assunto, discussão, tipo, aspecto, possível, exemplo, tempo, dificuldade, extremamente, comentário
24	**Judicial System - Constitution**	lei, constituição, legislação, público, projeto, tribunal, jurídico, federal, direito, decisão, constitucional, competência, legal, executivo, penal, razão, próprio, estrangeirar, procedimento, ministério, supremo, juiz, judiciário, medido, estatuto, ordem, serviço, código, norma, órgão
25	**Education**	educação, universidade, brasileiro, conhecimento, médico, qualidade, ciência, escola, cultura, instituição, brasil, professor, formação, profissional, grande, cultural, público, estudante, superior, ano, paulo, experiência, unesco, bom, oportunidade, senhor, importante, instituto, tempo, ministério
26	**Middle East**	paz, guerra, país, conflito, israel, iraque, onu, arma, irã, palestino, sírio, nação, povo, brasil, contra, internacional, médio, segurança, resolução, situação, unido, posição, conselho, nuclear, árabe, líbano, mundo, região, território, próprio
27	**Itamaraty**	relação, brasil, exterior, repúblico, embaixada, embaixador, presidente, chefe, diplomático, diplomata, missão, ministério, carreira, ordem, secretário, função, cargo, branco, itamaraty, indicação, classe, permanente, divisão, conselheiro, instituto, brasileiro, relatório, mérito, federal, bilateral

28	**EU**	europeu, união, país, frança, europa, rússia, alemanha, político, grande, ano, relação, itália, francês, presidente, crise, alemão, ucrânia, russo, guerra, européia, população, grécia, eleição, século, soviético, italiano, maior, espanha, história, irlanda
29	**Meeting Protocol 2**	requerimento, senador, discussão, votação, senado, item, federal, comissão, autoria, regimentar, matéria, plenário, autor, providência, solicitação, mesa, extrapauta, urgência, incisar, deliberação, francisco, presidência, relatório, dornelles, aprovado, período, inclusão, seguinte, aprovação, missão
30	**Haiti**	brasileiro, brasil, país, situação, haiti, exterior, japão, comunidade, itamaraty, ministério, número, haitiano, imigrante, imigração, problema, autoridade, consulado, migração, japonês, cidadão, estrangeiro, consular, turismo, assistência, migratório, ano, dificuldade, condição, cidade, família
31	**Senator 2**	senador, palavra, suplicy, cristovam, eduardo, ana, amélia, buarque, aloysio, jorge, ferraço, viana, nunes, henrique, ricardo, vanessa, luiz, ordem, ferreiro, grazziotin, seguido, tasso, anastasia, lasier, jereissati, indagação, ex^{a}s, requião, raupp, monteiro
32	**Finance - BNDES**	ano, milhão, recurso, banco, bilião, valor, real, dólar, financeiro, financiamento, r, orçamentar, dívida, dinheiro, bndes, investimento, número, crédito, serviço, sistema, pagamento, último, operação, prazo, mês, cincar, longo, salário, central, situação

33	**Mercosul**	mercosul, país, brasil, argentino, sul, américa, integração, paraguai, relação, chile, uruguai, região, latinar, bloco, brasileiro, peru, importante, venezuela, equador, pacífico, colômbio, comercial, vizinho, maior, acordo, grande, comum, unasul, regional, pontar
34	**Technology - Aerospace**	tecnologia, projeto, avião, aéreo, defeso, brasileiro, aeronave, satélite, desenvolvimento, indústria, tecnológico, área, brasil, sistema, embraer, nacional, base, espacial, lançamento, equipamento, capacidade, transferência, aeronáutico, alcântara, empresa, importante, estratégico, projetos, industrial, ciência
35	**Meeting Protocol 3**	repúblico, relatório, item, embaixador, senado, federal, senador, exterior, cargo, relação, apreciação, mensagem, relator, relatoria, indicação, terminativo, autoria, brasil, classe, ministério, carreira, observação, nome, matéria, comissão, diplomata, leitura, ad, hoc, sf
36	**Unknown 3**	político, país, sociedade, social, importante, exemplo, pontar, política, internacional, sistema, desenvolvimento, grande, mundo, nacional, novo, fundamental, próprio, capacidade, visão, elemento, aspecto, papel, instituição, realidade, contexto, diferente, conceito, longo, estratégia, maneiro
37	**Navy - Submarine**	marinhar, nuclear, submarino, tecnologia, navio, construção, brasil, mar, brasileiro, projeto, ano, área, recurso, grande, naval, marítimo, base, estação, plataforma, antártica, água, projetos, instalação, capacidade, tempo, agência, material, ministério, energia, almirante

38 **Audience**	dr, palavra, professor, paulo, prof, presença, audiência, mesa, convidado, josé, embaixada, público, exposição, representante, diretor, painel, carlos, secretário, minuto, conselheiro, universidade, departamento, roberto, assessor, debate, luiz, palestrantes, tempo, jorge, noite
39 **Cooperation**	brasil, país, cooperação, área, importante, relação, brasileiro, ano, bilateral, grande, maior, diálogo, importância, relacionamento, parceria, parceiro, oportunidade, inclusive, comercial, internacional, intercâmbio, desenvolvimento, novo, presença, investimento, comércio, projetos, experiência, cultural, atuação
40 **Deputy**	deputado, palavra, pergunta, carlos, s.exa, obrigado, sr, conosco, presença, congresso, dr, nobre, mara, riso, nelson, cão, autor, v.sa, rosa, raul, audiência, colega, ivan, seguido, legislativo, presidência, jungmann, antonio, questionamento, hauly

B.3 Terms when $k = 50$

Table 15: Politician Speech Topics ($k = 50$)

Topic		Keys (Top 30)
1	**Climate Change**	desenvolvimento, país, conferência, brasil, sustentável, ambiental, mudança, questão, nação, planeta, mundo, clima, discussão, global, água, importante, climático, grande, ano, meta, objetivos, biodiversidade, maior, mundial, internacional, unido, fórum, emissão, compromisso, sociedade
2	**Venezuela - Unasul**	venezuela, presidente, paraguai, colômbio, brasileiro, democrático, país, oposição, democracia, venezuelano, político, eleição, situação, brasil, paraguaio, chávez, manifestação, oea, decisão, unasul, contra, relação, farc, posição, chanceler, golpe, diálogo, colombiano, eleitoral, lado
3	**Human Rights**	direito, humano, mulher, brasil, criança, social, país, trabalhador, sociedade, convenção, importante, proteção, questão, atenção, civil, condição, internacional, defensor, violação, ano, escravo, contra, igualdade, morte, oit, família, situação, nacional, liberdade, caso
4	**Army - Defense**	defeso, força, armada, exército, nacional, comandante, general, aeronáutico, ministério, estratégico, marinhar, guerra, operação, brasileiro, segurança, missão, civil, oficial, soberania, presença, orçamentar, fronteira, situação, necessidade, estado-maior, tropa, condição, preocupação, estratégia, homem
5	**Exportation**	produto, brasil, exportação, brasileiro, grande, produção, agricultura, produtor, setor, indústria, comércio, país, agrícola, exportador, importação, maior, pequeno, preço, soja, carnar, café, área, sul, produtivo, exemplo, comercial, agricultor, valor, importante, alimento

6	**Diplomacy**	relação, brasil, exterior, presidente, chefe, diplomata, ministério, embaixada, república, diplomático, carreira, embaixador, função, secretário, missão, cargo, ordem, branco, classe, permanente, indicação, divisão, relatório, brasileiro, instituto, conselheiro, itamaraty, federal, mérito, internacional
7	**Technology - Aerospace**	aéreo, avião, aeronave, satélite, aeronáutico, projeto, embraer, lançamento, alcântara, espacial, aviação, sistema, civil, aeroporto, brasileiro, equipamento, brigadeiro, base, operação, veículo, ano, radar, anac, vôo, área, técnico, foguete, agência, controlador, sivam
8	**Crime - Police**	polícia, federal, público, segurança, crime, policial, tráfico, droga, fronteirar, arma, janeiro, problema, operação, paulo, ministério, civil, estadual, violência, investigação, criminoso, sociedade, ações, fronteira, estado, contra, exemplo, ilícito, ação, cidadão, dado
9	**Education**	educação, universidade, médico, escola, cultura, qualidade, brasil, instituição, grande, professor, profissional, ano, paulo, cultural, brasileiro, estudante, formação, conhecimento, público, jovem, unesco, superior, cidade, curso, experiência, bom, ciência, oportunidade, reconhecimento, aluno
10	**Finance**	empresa, banco, financeiro, dívida, setor, financiamento, serviço, varig, bndes, fiscal, crédito, capital, brasil, solução, grande, investimento, público, país, crise, dinheiro, central, tributário, operação, valor, juro, situação, sistema, problema, real, pagamento

11	**Trade Negotiation**	negociação, comércio, país, brasil, omc, alca, acordo, união, comercial, área, européia, importante, serviço, â, subsídio, rodado, mercosul, reunião, regra, agrícola, exemplo, acesso, tema, setor, desenvolvimento, mercado, agricultura, negociador, brasileiro, grande
12	**Energy - Bolivia**	energia, bolívia, brasil, petrobras, petróleo, gás, preço, brasileiro, itaipu, boliviano, produção, energético, paraguai, questão, grande, investimento, importante, usina, natural, condição, elétrica, momento, pontar, maior, país, contrato, gasoduto, seguinte, mina, exploração
13	**Meeting Protocol 1**	comissão, exterior, relação, nacional, defeso, reunião, senado, audiência, público, aberto, trabalho, ordinário, sessão, realização, anterior, subcomissão, leitura, legislativo, requerimento, federal, regimental, realizado, ministério, extraordinário, número, presidência, membro, aprovação, legislatura, aprovado
14	**Navy - Submarine**	nuclear, marinhar, submarino, navio, brasil, mar, construção, base, naval, tecnologia, marítimo, brasileiro, energia, grande, estação, plataforma, antártica, ano, instalação, área, projeto, agência, água, propulsão, material, almirante, capacidade, convencional, urânio, novo
15	**EU**	europeu, união, país, brasil, europa, frança, alemanha, grande, relação, itália, francês, político, européia, alemão, espanha, crise, maior, grécia, ano, importante, suíço, irlanda, interessante, italiano, euro, econômica, portugal, comum, pontar, presidente
16	**Senator 1**	senador, josé, palavra, costa, joão, relator, marcelo, jefferson, hélio, crivella, péres, arthur, mozarildo, virgílio, capiberibe, gilberto, cavalcanti, gentileza, augusto, sentado, saturnino, agripino, favorável, roberto, requerimento, votação, viegas, tião, gomar, sarney

17	**Ambassadors**	embaixador, celso, amorim, sérgio, exposição, palavra, luiz, vieira, josé, itamaraty, s[a], alberto, roberto, presença, souza, guimarães, mauro, ambos, maria, barbosa, rubens, pinheiro, figueiredo, antônio, carlos, nome, amaral, machadar, samuel, antonio
18	**Technology**	tecnologia, defeso, desenvolvimento, indústria, nacional, área, brasil, tecnológico, brasileiro, país, estratégico, empresa, ciência, importante, industrial, capacidade, projetos, inovação, investimento, exemplo, estratégia, grande, setor, sistema, transferência, base, produto, produção, parceria, programa
19	**Africa - CLP**	país, áfrico, brasil, presidente, português, língua, grande, sul, angola, relação, africanar, brasileiro, comunidade, independência, religioso, igreja, timor, africano, moçambique, ano, portugal, importante, católico, continente, santo, população, religião, guiné, paz, presença
20	**Numbers**	ano, milhão, bilião, us, dólar, país, número, pib, maior, crescimento, cincar, investimento, período, último, dado, população, brasil, médio, valor, exportação, economia, dez, total, principal, habitante, alto, r, próximo, seis, mês
21	**Senator 2**	senador, eduardo, palavra, suplicy, pedrar, luiz, simon, henrique, azeredo, francisco, exa, ribeiro, tuma, dornelles, romeu, flexa, heráclito, seguido, paulo, miranda, santo, cyro, ex[a]s, renan, mão, geraldo, calheiros, alvaro, bernardo, chave
22	**Russia**	presidente, rússia, eleição, relação, político, país, trump, russo, ucrânia, partido, questão, guerra, norte, eleitoral, ano, novo, soviético, eleito, união, maioria, momento, território, paquistão, fim, líder, coreia, campanha, grande, voto, europa

23	**Senator 3**	senador, cristovam, ana, amélia, buarque, jorge, aloysio, ferraço, viana, palavra, ricardo, nunes, vanessa, ferreiro, grazziotin, anastasia, requião, tasso, lasier, jereissati, roberto, monteiro, raupp, seguido, martim, arruda, inácio, jarbas, bezerro, lobão
24	**Unknown 1**	internacional, político, país, sociedade, importante, social, mundo, exemplo, novo, pontar, política, sistema, próprio, elemento, papel, contexto, grande, desenvolvimento, fundamental, capacidade, organização, diferente, maneiro, instituição, aspecto, tema, dimensão, visão, cenário, global
25	**Job Titles**	dr, professor, prof, paulo, embaixada, embaixador, presença, painel, palavra, josé, secretário, diretor, universidade, conselheiro, repúblico, audiência, convidado, noite, ministério, departamento, carlos, representante, brasília, palestrantes, relação, jorge, próximo, chefe, federação, assessor
26	**Meeting Protocol 2**	repúblico, embaixador, federal, relatório, senado, cargo, mensagem, item, apreciação, indicação, brasil, ministério, relação, exterior, carreira, diplomata, classe, nome, relator, relatoria, senador, observação, presidente, coletiva, autoria, sf, leitura, terminativo, cumulativamente, indicado
27	**Unknown 2**	questão, relação, momento, fato, problema, importante, claro, situação, pontar, maneiro, evidentemente, inclusive, posição, nenhum, exatamente, preocupação, próprio, discussão, dúvida, verdade, algum, exemplo, tipo, aspecto, possível, coisa, extremamente, possibilidade, função, assunto
28	**Conflicts - War**	povo, guerra, cubar, contra, iraque, país, mundo, cubano, arma, momento, história, fato, liberdade, posição, homem, democracia, manifestação, nenhum, verdade, episódio, paz, morte, razão, ditadura, solidariedade, regime, inclusive, opinião, conflito, contrário

29	**Unknown 3**	gente, problema, mundo, bom, grande, coisa, ano, tempo, nenhum, ninguém, brasil, pontar, exemplo, alguém, dinheiro, quê, daqui, difícil, maneiro, verdade, lado, época, lugar, diferente, história, pessoal, realidade, claro, interessante, nisso
30	**China - BRICs**	chinar, brasil, país, mundo, economia, político, grande, chinês, ano, crescimento, mundial, internacional, brasileiro, crise, relação, brics, estado, unido, comércio, econômica, comercial, índio, econômico, último, maior, investimento, indústria, novo, global, europa
31	**UN**	nação, conselho, onu, haiti, paz, segurança, brasileiro, unido, brasil, país, missão, internacional, membro, resolução, organização, situação, decisão, presidente, permanente, tropa, presença, geral, desenvolvimento, participação, civil, ano, diálogo, representante, junho, contingente
32	**Meeting Protocol 3**	comissão, assunto, presidente, inclusive, sugestão, convite, presidência, possível, próximo, iniciativo, membro, conhecimento, esclarecimento, disposição, ministro, manifestação, convocação, viagem, requerimento, diálogo, propósito, entendimento, solicitação, objeto, oportunidade, outubro, bom, setembro, plenário, representante
33	**Mercosul**	mercosul, país, brasil, sul, argentino, américa, integração, relação, importante, chile, latinar, bloco, região, uruguai, grande, peru, pacífico, maior, paraguai, comercial, acordo, vizinho, comércio, equador, brasileiro, regional, pontar, comum, continente, aliançar
34	**Meeting Protocol 4**	deputado, câmara, audiência, mesa, comissão, público, representante, sr, presença, senhor, nome, seminário, presidente, colega, palma, congresso, convidado, momento, oportunidade, convite, evento, exmo, frente, obrigado, bom, debate, iniciativo, importância, nacional, início

35	**Meeting Protocol 5**	ordem, palavra, intervenção, pergunta, favor, microfone, tempo, riso, consideração, algum, questionamento, obrigado, respostar, indagação, pronunciamento, final, resposta, comentário, inaudível, exatamente, seguido, desculpe-me, minutar, orador, permita-me, exposição, observação, formulado, esclarecimento, interrupção
36	**Middle East**	país, israel, irã, árabe, sírio, palestino, médio, político, região, líbio, paz, internacional, conflito, grande, egito, novo, islâmico, regime, líbano, ano, mundo, turquia, unido, saúde, tunísia, arábio, contra, conselho, segurança, muçulmano
37	**Congress**	internacional, legislativo, texto, cooperação, nacional, projeto, repúblico, aprovação, matéria, convenção, federativo, congresso, brasil, câmara, análise, relação, parte, presidente, técnico, acordo, tratado, jurídico, artigo, protocolo, vigor, organização, âmbito, relativo, brasília, oportuno
38	**Judicial System - Constitution**	lei, constituição, legislação, tribunal, projeto, público, federal, jurídico, direito, constitucional, decisão, competência, nacional, legal, seguinte, executivo, supremo, norma, medido, juiz, próprio, princípio, social, tal, procedimento, estrangeirar, alteração, razão, judiciário, código
39	**Information - Spy**	informação, dado, comunicação, internet, redar, brasileiro, acesso, empresa, sistema, serviço, questão, cidadão, segurança, público, tipo, conhecimento, espionagem, brasil, notícia, agência, autoridade, importante, nenhum, algum, conteúdo, civil, nome, satélite, sigilar, inclusive
40	**Brazilian Presidents**	presidente, lula, fernando, colega, collor, momento, importante, oportunidade, dilma, inclusive, henrique, mim, certeza, líder, franco, caro, cardoso, senado, tempo, expresidente, governador, pt, vice-presidente, companheiro, razão, naquela, dúvida, bom, viagem, repúblico

41	**Cooperation**	brasil, país, cooperação, área, relação, importante, brasileiro, bilateral, comércio, grande, comercial, relacionamento, embaixada, desenvolvimento, investimento, maior, intercâmbio, presença, inclusive, parceiro, oportunidade, turismo, possibilidade, ano, cultural, diálogo, parceria, novo, potencial, principal
42	**Terrorism**	inteligência, segurança, grande, evento, atividade, órgão, área, brasileiro, terrorismo, nacional, atividades, agência, público, país, abin, sistema, copar, informação, serviço, terrorista, ações, organização, mundo, tipo, institucional, conhecimento, coordenação, sociedade, tempo, nível
43	**Itamaraty**	político, brasileiro, brasil, relação, país, diplomacia, internacional, itamaraty, posição, novo, opinião, exterior, papel, sociedade, nacional, política, mundo, compromisso, nação, fundamental, bom, segurança, ideológico, diplomático, fato, direito, presidente, avaliação, grande, necessidade
44	**Budget**	projeto, recurso, ministério, projetos, orçamentar, ano, nacional, emenda, necessidade, ações, valor, congresso, ação, prioridade, planejamento, necessário, r, real, execução, importante, prazo, programa, orçamentária, obra, órgão, implantação, dificuldade, apresentado, geral, gestão
45	**North America (USA, Canada, Mexico)**	estado, unido, americano, brasil, relação, presidente, méxico, país, canadá, brasileiro, américa, congresso, mundo, maior, norte-americano, importante, washington, mexicano, problema, norte-americana, américas, obama, americana, forte, novo, exemplo, inclusive, lado, nação, bush
46	**Immigration - Japan**	brasileiro, brasil, exterior, país, situação, ministério, japão, comunidade, imigrante, número, imigração, itamaraty, autoridade, migração, estrangeiro, consulado, japonês, relação, cidadão, família, consular, assistência, problema, haitiano, turismo, migratório, condição, dificuldade, entrado, turista

47	**Meeting Protocol 6**	reunião, hora, minuto, encerrado, presença, próximo, sessão, plenário, conosco, secreto, congresso, deputado, pergunta, mara, legislativo, cão, iniciado, comissão, nacional, portal, horário, legislatura, trabalho, deliberativo, audiência, df, reaberto, manhã, boletim, cnpj
48	**Meeting Protocol 7**	requerimento, item, senador, discussão, votação, senado, relatório, matéria, autoria, regimentar, federal, projeto, relator, plenário, terminativo, aprovação, relatoria, extrapauta, providência, deliberação, leitura, comissão, incisar, mesa, aprovado, urgência, ad, solicitação, inversão, inclusão
49	**Amazon**	amazônia, região, área, fronteirar, indígena, terra, índio, roraima, território, problema, questão, norte, guiana, comunidade, população, município, faixar, amazônica, amazona, grande, quilômetros, sul, cidade, grosso, acre, povo, água, nacional, manaus, soberania
50	**Deputy**	palavra, dr, carlos, deputado, s.exa, autor, antonio, minuto, fernando, nobre, requerimento, exposição, gabeira, luiz, v.sa, rosa, josé, raul, ivan, paulo, presidência, mourão, jungmann, hauly, joão, seguido, nilson, expositor, pannunzio, valente

B.4 Terms when $k = 60$ – Chosen model

Table 16: Politician Speech Topics ($k = 60$)

	Topic	Keys (Top 30)
1	**Position - Attitude**	político, brasileiro, posição, itamaraty, fato, relação, opinião, nenhum, claro, contrário, razão, atitude, episódio, jornal, contra, manifestação, preocupação, momento, lula, ideológico, assunto, maneiro, posturar, notícia, soberania, verdade, situação, jornalista, diferente, evidente
2	**Climate Change**	desenvolvimento, conferência, país, ambiental, sustentável, mudança, brasil, planeta, nação, clima, ano, questão, global, água, internacional, discussão, climático, protocolo, biodiversidade, grande, emissão, unido, objetivos, mundo, compromisso, meta, fórum, mundial, maior, convenção
3	**Diplomacy**	diplomacia, brasil, brasileiro, político, internacional, relação, país, novo, exterior, embaixador, nação, opinião, sociedade, tempo, nacional, professor, necessidade, presidente, compromisso, segurança, dr, conselho, audiência, nome, ministério, organização, ex-ministro, direito, unido, mundo
4	**National Borders**	região, fronteirar, grande, sul, cidade, norte, projeto, santo, problema, grosso, acre, quilômetros, município, faixar, infraestrutura, peru, obra, fronteira, pontar, lado, construção, principalmente, porto, prefeito, ferrovia, naquela, nordeste, logístico, panamá, maior
5	**Meeting Protocol 1**	reunião, hora, minuto, encerrado, presença, próximo, plenário, sessão, pergunta, secreto, deputado, congresso, conosco, comissão, iniciado, mara, cão, legislativo, nacional, legislatura, portal, horário, df, manhã, boletim, cnpj, acessibilidade, english, deliberativo, 55âª

6	**House of Representantives Committee**	deputado, audiência, público, mesa, câmara, presença, representante, comissão, sr, convidado, exposição, palma, seminário, tempo, evento, exmo, convite, senhor, trabalho, nome, debate, minuto, início, tv, eduardo, realização, participação, bom, almeida, nelson
7	**Multilateral Negotiation**	negociação, comércio, país, união, acordo, omc, alca, brasil, comercial, européia, área, â, serviço, mercosul, rodado, subsídio, agrícola, regra, importante, europeu, tema, reunião, produto, acesso, desenvolvimento, agricultura, setor, negociador, mercado, multilateral
8	**Worker's Condition - ILO**	social, mulher, trabalhador, criança, ano, família, previdência, condição, serviço, local, país, situação, atenção, salário, escravo, tempo, direito, oit, lei, número, legislação, sexual, relação, mãe, população, menino, morte, acesso, obrigado, público
9	**Intelligence - Security**	informação, inteligência, dado, sistema, segurança, internet, agência, brasileiro, comunicação, redar, serviço, abin, empresa, atividade, acesso, tipo, espionagem, cidadão, conhecimento, cibernético, legislação, público, autoridade, órgão, proteção, nacional, conteúdo, lei, nome, algum
10	**Unknown 1**	questão, discussão, exatamente, inclusive, importante, pontar, relação, claro, preocupação, fundamental, problema, próprio, verdade, evidentemente, fato, dificuldade, assunto, visão, momento, dúvida, colocado, função, principalmente, extremamente, algum, ponto, posição, comentário, papel, tempo
11	**Crime - Police**	polícia, federal, público, segurança, crime, policial, tráfico, droga, janeiro, operação, evento, grande, órgão, ministério, civil, copar, estadual, ações, lei, paulo, penal, investigação, criminoso, inteligência, ilícito, fronteira, legislação, atuação, judiciário, prevenção

12	**Budget**	ano, milhão, recurso, bilião, orçamentar, valor, real, r, número, cincar, dólar, projeto, último, emenda, dez, prazo, mês, dado, maior, seis, próximo, orçamentária, oito, período, longo, investimento, total, mínimo, ministério, despesa
13	**Brazilian Image**	brasil, país, importante, brasileiro, grande, mundo, relação, maior, papel, exemplo, importância, extremamente, pontar, bom, inclusive, rico, posição, experiência, presença, nesses, enorme, pequeno, fundamental, forte, mundial, através, tamanho, contribuição, momento, oportunidade
14	**UN - Security Council**	paz, onu, haiti, nação, israel, iraque, conselho, segurança, palestino, guerra, conflito, missão, unido, resolução, situação, povo, brasil, tropa, internacional, líbano, país, território, membro, solução, presidente, lado, população, iraquiano, iniciativo, momento
15	**EU**	europeu, união, europa, país, frança, rússia, alemanha, relação, grande, itália, francês, ano, alemão, européia, político, russo, ucrânia, guerra, soviético, presidente, grécia, suíço, irlanda, república, italiano, população, sérvio, império, norte, inglaterra
16	**Senate Committee**	relatório, item, república, relator, senado, federal, matéria, mensagem, autoria, relatoria, apreciação, terminativo, senador, cargo, discussão, projeto, leitura, exterior, brasil, ad, indicação, diplomata, carreira, classe, hoc, ministério, observação, sf, coletiva, dezembro
17	**Navy - Submarine**	nuclear, marinhar, submarino, navio, mar, construção, tecnologia, naval, brasileiro, marítimo, energia, brasil, base, estação, área, antártica, plataforma, água, recurso, projeto, propulsão, almirante, instalação, material, convencional, urânio, capacidade, agência, ciência, combustível

18	**Mercosul**	mercosul, argentino, américa, integração, país, paraguai, sul, brasil, relação, chile, uruguai, bloco, latinar, venezuela, equador, região, paraguaio, peru, importante, comum, pacífico, acordo, comercial, vizinho, unasul, regional, bolívia, aliançar, colômbio, brasileiro
19	**Alcantara - Aerospace**	satélite, projeto, base, espacial, lançamento, alcântara, brasileiro, área, tecnologia, comunicação, veículo, ucrânia, recurso, foguete, agência, técnico, informação, capacidade, bandar, ciência, nenhum, soberania, inclusive, tecnológico, comunidade, desenvolvimento, lançador, possibilidade, conhecimento, próprio
20	**Ambassadors**	assunto, celso, presidente, amorim, inclusive, sugestão, comissão, próximo, convite, diálogo, josé, ministro, convocação, esclarecimento, entendimento, possível, informação, oportunidade, pinheiro, iniciativo, paulo, heráclito, ambos, diverso, guimarães, próprio, samuel, propósito, disposição, objeto
21	**Defense**	defeso, força, armada, exército, comandante, nacional, general, ministério, aeronáutico, marinhar, estratégico, operação, guerra, orçamentar, segurança, lei, missão, necessidade, oficial, projetos, presença, brasileiro, situação, soberania, cibernético, fronteira, maior, estado-maior, tropa, civil
22	**Defense Technology - Embraer**	defeso, tecnologia, indústria, desenvolvimento, projeto, nacional, tecnológico, projetos, estratégico, brasileiro, embraer, avião, área, sistema, capacidade, empresa, transferência, equipamento, industrial, aeronave, aéreo, inovação, importante, estratégia, brasil, aeronáutico, produto, produção, exemplo, ciência

23	**Cooperation**	cooperação, brasil, área, relação, país, bilateral, brasileiro, importante, internacional, cultural, ano, relacionamento, desenvolvimento, diálogo, intercâmbio, parceria, técnico, projetos, maior, embaixada, turismo, comercial, novo, oportunidade, organização, missão, atuação, importância, parceiro, entendimento
24	**Agricultural**	agricultura, produtor, grande, produção, brasil, produto, agrícola, pequeno, área, soja, preço, setor, brasileiro, café, subsídio, agricultor, rural, sul, maior, carnar, embrapa, tabaco, algodão, alimento, novo, indústria, importante, tonelada, bom, familiar
25	**Amazon**	amazônia, área, indígena, terra, índio, região, roraima, guiana, território, fronteirar, amazônica, comunidade, água, povo, população, florestar, federal, funai, amapá, soberania, amazona, problema, brasileiro, sol, demarcação, questão, suriname, manaus, nacional, bom
26	**Immigration - Japan**	brasileiro, brasil, exterior, país, situação, japão, comunidade, imigrante, imigração, migração, estrangeiro, ministério, consulado, autoridade, cidadão, número, itamaraty, japonês, consular, turismo, portugal, haitiano, assistência, migratório, turista, família, ano, espanha, entrado, refugiado
27	**Senator 1**	senador, eduardo, suplicy, cristovam, palavra, ana, amélia, buarque, luiz, henrique, vanessa, francisco, grazziotin, dornelles, seguido, ex^{a}s, rodrigar, azeredo, randolfe, inácio, miranda, cyro, arruda, indagação, anibal, diniz, marta, pds, formulado, braga
28	**Religion**	povo, homem, história, brasileiro, religioso, deus, santo, igreja, negro, paulo, cultura, católico, liberdade, religião, nome, irmão, família, ano, companheiro, grande, contra, homenagem, mão, pai, verdade, humanidade, momento, solidariedade, democracia, mundo

29	**Jobs Title**	dr, prof, paulo, palavra, presença, professor, embaixada, josé, secretário, conselheiro, diretor, painel, roberto, carlos, barbosa, convidado, república, embaixador, departamento, v.sa, ministério, antônio, representante, universidade, audiência, márcio, luiz, palestrantes, rubens, noite
30	**Unknown 2**	nacional, sociedade, ministério, área, brasileiro, ações, civil, órgão, participação, político, importante, política, sistema, ação, atividades, conselho, público, objetivo, instituição, gestão, organização, social, necessidade, âmbito, desenvolvimento, atuação, executivo, recurso, específico, necessário
31	**Ambassador appointment**	embaixador, república, sérgio, palavra, maria, exposição, indicado, luiz, souza, vieira, antonio, josé, missão, patriota, indicação, nome, alberto, cargo, exa, voto, lúcia, presença, ambos, machadar, figueiredo, marco, cumulativamente, mello, aberto, mauro
32	**Terrorism**	arma, colômbio, país, terrorismo, contra, brasil, terrorista, presidente, brasileiro, guerra, farc, colombiano, grupo, bomba, munição, paquistão, organização, relação, armamento, desarmamento, fogo, guerrilhar, conflito, assunto, negociação, momento, narcotráfico, morte, violência, tempo
33	**Unknown 3**	presidente, oportunidade, colega, tempo, certeza, senhor, bom, caro, importância, s[a], mina, importante, iniciativo, governador, frente, nome, geral, dúvida, parabém, conhecimento, experiência, alegria, companheiro, obrigado, feliz, exposição, lugar, desculpa, exatamente, brilhante
34	**Brazilian Presidents**	presidente, lula, fernando, cubar, cubano, collor, henrique, dilma, república, inclusive, cardoso, franco, viagem, ex-presidente, vice-presidente, época, líder, pt, senado, presidência, itamar, rousseff, mim, ministro, sarney, ocasião, inácio, cpi, fidel, sr

35	**Unknown 4**	político, internacional, mundo, país, novo, grande, ano, mundial, social, visão, pontar, global, econômica, cenário, política, exemplo, econômico, sociedade, desenvolvimento, economia, papel, década, século, estratégico, capacidade, fundamental, importante, longo, sistema, elemento
36	**Unknown 5**	crise, brasil, país, político, ano, brasileiro, internacional, medida, economia, relação, mundo, próximo, público, econômica, mundial, euro, europeu, novo, problema, estado, europa, fim, nacional, unido, crescimento, prazo, momento, situação, embaixador, comissão
37	**Meeting Protocol 2**	brasileiro, senador, comissão, país, informação, nação, brasil, novo, unido, reunião, exa, nacional, junho, internacional, embaixador, presidente, desenvolvimento, bolívia, presença, final, evento, político, questão, presidência, o, relação, boliviano, sírio, próximo, delegação
38	**Unknown 6**	gente, problema, bom, coisa, mundo, ano, tempo, ninguém, exemplo, grande, dinheiro, quê, nenhum, alguém, pontar, difícil, lado, verdade, daqui, nisso, lugar, maneiro, diferente, época, interessante, cima, frente, algum, negócio, caro
39	**Senator 2**	senador, jorge, viana, aloysio, ferraço, roberto, ricardo, nunes, ferreiro, palavra, flexa, tuma, josé, tasso, requião, ribeiro, agripino, anastasia, romeu, lasier, jereissati, relator, raupp, monteiro, martim, saturnino, bezerro, valdir, lobão, jarbas
40	**Arab countries**	país, irã, árabe, sírio, político, médio, líbio, região, egito, islâmico, ano, grande, regime, relação, internacional, saúde, conselho, tunísia, arábio, mundo, turquia, muçulmano, contra, iraniano, novo, segurança, potência, emirado, líder, sanção

41	**Education**	educação, universidade, escola, médico, professor, qualidade, ciência, instituição, estudante, público, superior, formação, conhecimento, brasileiro, ano, profissional, instituto, paulo, jovem, curso, reconhecimento, grande, aluno, estudo, federal, medicinar, avaliação, associação, fundação, cultura
42	**Africa**	país, áfrico, português, presidente, língua, sul, ano, brasil, angola, população, comunidade, africanar, grande, independência, embaixada, portugal, timor, africano, moçambique, costa, relação, continente, presença, guiné, brasileiro, nação, milhão, indonésio, congo, cabo
43	**Meeting Protocol 3**	requerimento, senado, votação, federal, discussão, comissão, regimentar, senador, item, plenário, autoria, matéria, solicitação, autor, aprovado, urgência, extrapauta, período, incisar, missão, seguinte, deliberação, inclusão, providência, convite, audiência, presidência, aprovação, regimental, realização
44	**Meeting Protocol 4**	comissão, relação, exterior, nacional, defeso, ministério, senado, presidente, congresso, câmara, membro, assunto, misto, plenário, representante, perante, presidência, executivo, apresentado, âmbito, encaminhado, sugestão, assessoria, econômicos, importância, relevância, realização, secretariar, objetivo, satisfação
45	**Human Rights**	direito, humano, internacional, convenção, organização, princípio, proteção, nação, liberdade, tribunal, civil, violação, conselho, contra, defensor, declaração, onu, próprio, oea, âmbito, sistema, brasil, conferência, obrigação, membro, igualdade, organismo, democrático, órgão, cidadão

46	**Itamaraty**	brasil, relação, exterior, repúblico, diplomata, carreira, embaixador, chefe, diplomático, embaixada, ministério, cargo, missão, presidente, classe, secretário, função, indicação, branco, permanente, ordem, relatório, federal, divisão, itamaraty, conselheiro, instituto, mérito, senado, delegação
47	**Senator 3**	senador, pedrar, palavra, simon, costa, marcelo, joão, jefferson, hélio, relator, crivella, péres, arthur, josé, virgílio, azeredo, sentado, capiberibe, favorável, votação, mozarildo, gilberto, cavalcanti, maciel, gentileza, eduardo, tourinho, mestrinho, peres, augusto
48	**Meeting Protocol 5**	reunião, comissão, aberto, exterior, defeso, sessão, senador, senado, trabalho, ordinário, extraordinário, leitura, audiência, anterior, subcomissão, presidência, público, nacional, realizado, relação, legislativo, número, aprovado, regimental, permanente, membro, integrante, próximo, presidente, ofício
49	**North America (USA, Canada, Mexico)**	estado, unido, americano, relação, méxico, brasil, canadá, américa, presidente, mundo, trump, norte-americano, congresso, obama, washington, novo, contra, mexicano, norte-americana, nação, américas, bush, americana, guerra, norte-americanos, comum, inclusive, latinar, norte, canadense
50	**Aviation**	aéreo, avião, civil, aviação, aeronave, varig, aeroporto, problema, sistema, aeronáutico, solução, empresa, companhia, vôo, anac, brigadeiro, responsabilidade, segurança, operação, setor, situação, sivam, controlador, tráfego, passageiro, equipamento, infraero, radar, brasília, condição
51	**Energy**	energia, bolívia, petrobras, petróleo, gás, brasil, preço, itaipu, produção, energético, boliviano, brasileiro, natural, investimento, usina, milhão, dólar, elétrica, contrato, grande, exploração, gasoduto, ano, condição, paraguai, matriz, fonte, questão, situação, geração

52	**Finance - BNDES**	banco, empresa, financeiro, financiamento, dívida, recurso, público, bndes, crédito, fiscal, capital, serviço, investimento, sistema, conta, central, tributário, operação, dinheiro, real, juro, pagamento, projetos, país, novo, valor, internacional, mecanismo, preço, próprio
53	**Deputy**	palavra, carlos, deputado, s.exa, dr, antonio, nobre, autor, luiz, gabeira, fernando, raul, rosa, seguido, jungmann, ivan, presidência, joão, mourão, hauly, minuto, nilson, pannunzio, josé, nelson, william, mendes, valente, fernandes, zarattini
54	**Unknown 7**	problema, maneiro, situação, exemplo, fato, próprio, grande, aspecto, importante, tipo, pontar, inclusive, possível, claro, tempo, evidentemente, momento, realidade, difícil, algum, coisa, diferente, preocupação, específico, modo, dúvida, dificuldade, verdade, bom, etc
55	**Unknown 8**	momento, presidente, decisão, nenhum, questão, medido, naquele, algum, final, possível, função, mão, tempo, verdade, fato, dúvida, tomado, base, provisório, respostar, necessário, conhecimento, certeza, modo, congresso, lado, consideração, ninguém, mim, seguinte
56	**Law - Constitution**	lei, projeto, legislativo, texto, nacional, constituição, aprovação, internacional, matéria, jurídico, repúblico, análise, congresso, câmara, artigo, constitucional, legislação, parte, executivo, federativo, presidente, convenção, técnico, federal, proposição, tratado, norma, vigor, 1º, acordo
57	**Exportation**	exportação, comércio, produto, brasil, us, ano, brasileiro, bilião, economia, investimento, país, comercial, indústria, empresa, setor, crescimento, pib, milhão, grande, maior, importação, dólar, industrial, desenvolvimento, principal, exportador, área, econômico, alto, valor

58	**Meeting Dialogues**	palavra, ordem, intervenção, microfone, favor, pergunta, riso, pronunciamento, resposta, consideração, indagação, inaudível, respostar, obrigado, final, questionamento, seguido, desculpe-me, algum, comentário, tradução, permita-me, inscrição, minutar, formulado, orador, ininteligível, interrupção, perdão, ah
59	**BRICs**	chinar, brasil, chinês, relação, índio, grande, país, brics, político, mundo, sul, embaixador, japão, economia, ano, coreia, ásia, rússia, coréia, presidente, unido, maior, asiático, presença, áfrico, problema, último, fato, parceiro, vietnã
60	**Venezuela**	venezuela, presidente, eleição, democrático, democracia, político, oposição, eleitoral, venezuelano, partido, situação, regime, chávez, eleito, contra, povo, constituição, líder, país, inclusive, liberdade, manifestação, golpe, popular, ditadura, mandatar, hugo, nacional, voto, oea