UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Câmpus de Presidente Prudente

MARIA EDUARDA DO PRADO E QUEIROZ

Uma Análise dos Tweets de Pré-Candidatos a Presidência do Brasil: Aplicação do Algoritmo de *Latent Dirichlet Allocation* (LDA)

PRESIDENTE PRUDENTE
2021/2022

MARIA EDUARDA DO PRADO E QUEIROZ

Uma Análise dos Tweets de Pré-Candidatos a Presidência do Brasil: Aplicação do Algoritmo de *Latent Dirichlet Allocation* (LDA)

Relatório Final para Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Estatística da FCT/Unesp para aproveitamento na disciplina Trabalho de Conclusão de Curso.

Orientador: Prof. Klaus Schlünzen Junior.

PRESIDENTE PRUDENTE

2021/2022

Q3a

Queiroz, Maria Eduarda do Prado e
Uma Análise dos Tweets de Pré-Candidatos a Presidência do Brasil : Aplicação do Algoritmo de Latent Dirichlet Allocation (LDA) / Maria Eduarda do Prado e Queiroz. -- Presidente Prudente, 2022
60 p. : il., tabs.

Trabalho de conclusão de curso (Bacharelado - Estatística) - Universidade Estadual Paulista (Unesp), Faculdade de Ciências e Tecnologia, Presidente Prudente
Orientador: Klaus Schlünzen

1. Mineração de Texto. 2. Latent Dirichilet Allocation. 3. Modelos de Tópicos. 4. TF-IDF. 5. Análise de Sentimentos. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca da Faculdade de Ciências e Tecnologia, Presidente Prudente. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

TERMO DE APROVAÇÃO

MARIA EDUARDA DO PRADO E QUEIROZ

Uma Análise dos Tweets de Pré-Candidatos a Presidência do Brasil: Aplicação do Algoritmo de *Latent Dirichlet Allocation* (LDA)

Relatório de Final de Trabalho de Conclusão de Curso aprovado como requisito para obtenção de créditos na disciplina Trabalho de Conclusão do curso de graduação em Estatística da Faculdade de Ciências e Tecnologia da Unesp, pela seguinte banca examinadora:

Orientador: _____

Prof. Dr. Klaus Schlünzen Junior.
Departamento de Estatística

_____

Profa. Dra. Miriam Rodrigues Silvestre
Departamento de Estatística

_____

Prof. Dr. Fernando Antonio Moala
Departamento de Estatística

Presidente Prudente, 11 de Março de 2022.

Dedico este trabalho a meus amados pais Sandra e Cláudio que sempre me apoiaram e estiveram presentes nos momentos difíceis e ao meu namorado Marcelo que foi meu melhor amigo e companheiro durante toda a graduação.

RESUMO

O Latent Dirichlet Allocation (LDA) é um modelo generativo para grupos de dados discretos como corpus de texto. Modelos generativos são aqueles que aleatoriamente geram os dados a partir das variáveis latentes. Nesse estudo vamos analisar o perfil do Twitter de pré-candidatos a presidência do Brasil no ano de 2021, usando técnicas de mineração de texto como, frequência de termos através de nuvens de palavras, análise de sentimentos, análise de agrupamento e análise da frequência inversa do termo (TF-IDF), além da aplicação do algoritmo LDA para o grupo de perfis. Para a realização das análises foram colhidos oitenta mil tweets de cada pré-candidato no período de 08/09/2021 a 27/10/2021, os dados foram tratados eliminando qualquer caracter que fosse irrelevante para a análise. Os resultados indicaram os termos mais frequentes e relevantes para cada perfil e com a aplicação do algoritmo constatamos que existe um grande número de termos que compõem cada assunto e qual pré-candidato tem a maior probabilidade de ser citado em determinado assunto.

Palavras-chaves: Mineração de Texto. Latent Dirichilet Allocation. Modelos de Tópicos. TF-IDF. Análise de Sentimentos.

ABSTRACT

Latent Dirichlet Allocation (LDA) is a generative model for discrete data groups as a text corpus. Generative models are those that randomly generate data from latent variables. In this study we will analyze the Twitter profile of pre-candidates for the presidency of Brazil in the year 2021, using text mining techniques such as term frequency through word clouds, sentiment analysis, cluster analysis and inverse frequency analysis. of the term (TF-IDF), in addition to the application of the LDA algorithm for the profile group. In order to carry out the analysis, eighty thousand tweets were collected from each pre-candidate in the period from 09/08/2021 to 10/27/2021, the data were processed by eliminating any character that was irrelevant to the analysis. The results indicated the most frequent and relevant terms for each profile and with the application of the algorithm we found that there is a large number of terms that make up each subject and which pre-candidate is most likely to be cited in a given subject.

Keywords: Text Mining. Latent Dirichlet Allocation. Topic Models. TF-IDF. Sentiment Analysis.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE SIGLAS

TM - Text Mining

KDT - Knowledge Discovery frow Text (Descoberta de conhecimentos em Textos)

PLN - Processamento de Linguagem Natural

IA - Inteligência Artificial

TF - Frequência dos Termos

CLINK - Método da Ligação Completa

TF-IDF - Term Frequency – Inverse Document Frequency

IDF - Frequência Inversa do Termo

LDA - Latent Dirichlet Allocation

Sumário

1 Introdução

Sabemos que as relações sociais são de suma importância para a existência da humanidade. Ao passo da evolução, vemos uma crescente cada vez maior no desenvolvimento de tecnologias de informação e comunicação, tendo início no telefone, rádio e televisão, se estendendo hoje até as redes sociais e reuniões online, fazendo com que cada vez menos exista a necessidade de um encontro presencial. Grande parte desse progresso está diretamente ligada a Internet e seus benefícios. Ao falar de um período de 10 anos no Brasil, obtivemos um salto de 27 milhões de usuários de redes sociais em 2009, para 200 milhões em 2019, que nos mostra o quanto a população se tornou adepta a esses meios de comunicação.

Além das relações humanas propriamente ditas, os meios de comunicações também são frequentemente acessados como fontes de informações, de forma especial o Twitter, que quando usado da maneira correta é realmente capaz de nos trazer fontes verdadeiras de informações.

O Twitter é uma rede social criada em 2006 que rapidamente ganhou popularidade e em 2020 chegou a 1,3 bilhões de contas com cerca de 340 milhões de mensagens postadas por dia. A rede consiste em compartilhar mensagens curtas de até 280 caracteres, sendo esses textos de domínio público. Dado o número de usuários e a facilidade de se expressar, muitos políticos viram um meio de se comunicar de forma aberta e direta com a população, fazendo de suas contas um meio de comunicação oficial. No caso de alguns políticos o número de seguidores ultrapassa a casa dos milhões.

A *Text Mining* (TM) ou Mineração de texto se tornou possível com a aprendizagem de máquinas, devido à possibilidade do processamento de grande volume de dados. Estima-se que 85% de toda a informação do mundo estejam em forma de documentos textuais, ou seja, estruturas rígidas e complexas. Então um passo muito grande para a humanidade é, além de interpretar essas informações, poder quantificá-las e categorizá-las de forma que os conhecimentos provindos delas sejam melhor compreendidos. Nesse sentido, somos capazes de coletar os

textos postados em cada perfil e analisar o posicionamento político e como é sua interação com a sociedade por meio de técnicas de mineração de texto.

No ano de 2022 o Brasil viverá mais um período de eleições presidenciais. Diante desse cenário futuro, uma forma diferente de conhecer e avaliar os candidatos é usando a TM em suas contas do *Twitter*. Dessa forma poderemos verificar quais os assuntos mais comentados em cada perfil, o posicionamento político de cada candidato e, por fim, conseguir através de uma frase ou um pequeno conjunto de palavras, dizer qual candidato tem maior probabilidade de ser o autor.

Levando em conta as eleições presidências de 2022, esse estudo tem como objetivos específicos coletar dados das contas de alguns Pré-candidatos a presidência do Brasil, estudar e apresentar resultados sobre o perfil político de cada um usando as técnicas de mineração de texto, sendo o objetivo central a aplicação do modelo probabilístico de tópico LDA para prever qual pré-candidato está fortemente associado a quais assuntos.

Nesse estudo iremos apresentar como os dados são coletados e organizados, as teorias das técnicas para as análises dos dados coletados, por fim a descrição do modelo *Latent Dirichlet Allocation* e a estimação de seus parâmetros usando o Amostrador de Gibbs. Na próxima seção serão apresentadas as justificativas principais que levaram a construção desse estudo.

1.1 Justificativa

Ao acompanhar a evolução dos usuários de redes sociais em um período de 10 anos, percebe-se que houve uma migração do "MSN" para o Whatsapp e do Orkut para o Facebook e Twitter. Nesse mesmo período muitos Brasileiros passaram a ter acesso a internet, um salto foi dado de 27 milhões de usuários de redes sociais em 2009 para mais de 200 milhões em 2019. Com isso o acesso à informação, sendo ela verdadeira ou não, também cresceu e em consequência o número de pessoas que expõe suas opiniões de forma aberta nas redes sociais aumentou drasticamente.

Quando levamos os aspectos discutidos anteriormente em consideração percebemos que podemos colher dados de uma forma muito mais simples e sem elevados gastos de recursos financeiros, considerando para este estudo

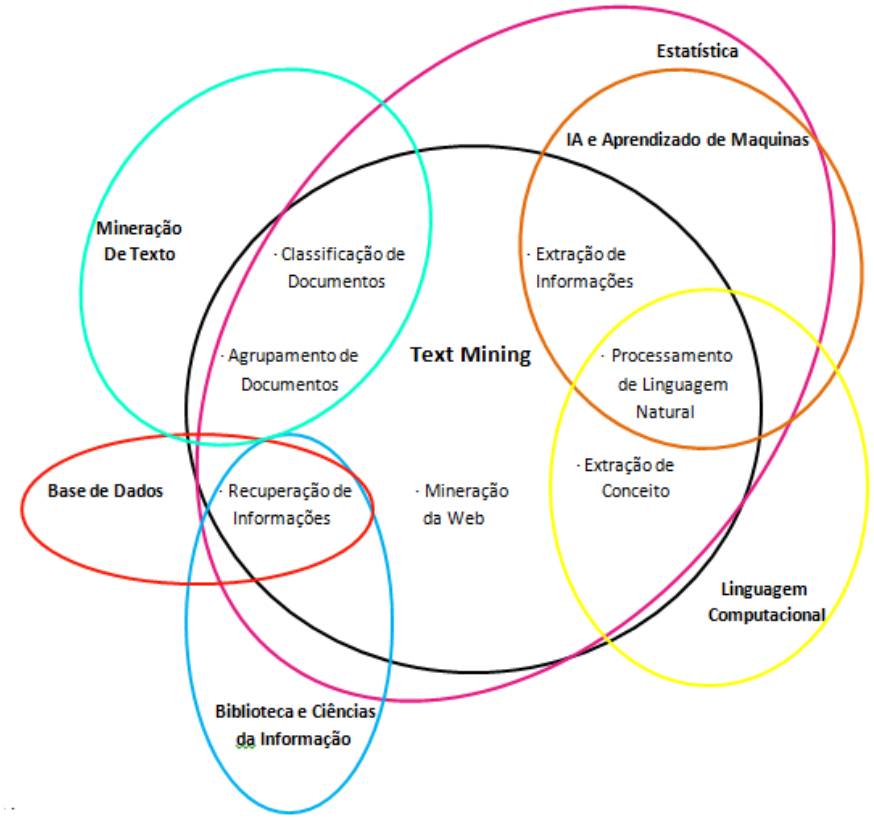
minidiscursos políticos que expressam uma opinião. Dessa forma, podemos colher esses minidiscursos no Twitter, pois a maioria das postagens de opinião envolvem textos limitados a 240 caracteres e estão disponíveis como dados públicos para a população. Logo aproveitamos dessa disponibilidade de informações a um custo bem reduzido para analisar e quantificar falas e discursos dos pré-candidatos.

2 Mineração de Texto (*Text Mining* - TM)

A Mineração de texto é um processo que utiliza algoritmos computacionais para extração de informações de dados não estruturados como arquivos de texto. Sullivan definiu a TM da seguinte forma "*Text mining* é o estudo e a prática de extrair informações de texto usando o princípio da linguística computacional" (SULLIVAN, 2004). Contudo a TM vai além de apenas compreender o que está escrito no texto e sim identificar padrões e associações quase que imperceptíveis ao homem sem à ajuda dos métodos computacionais (THURAISGHAM, 1999).

Um ponto importante para a *text mining* é analisar a informação relacionada ao objetivo de cada estudo, como por exemplo, uma empresa que deseja avaliar os comentários dos clientes a fim de desenvolver estratégias de marketing para auxiliar nas vendas direcionando a um público alvo ou melhorando a qualidade dos produtos, atendimento, etc. Entretanto a TM se relaciona com vários outros campos de estudo que englobam diversas áreas que vai desde o entendimento do problema até a obtenção de informações valiosas. É possível visualizar esses campos e conexões por meio do Diagrama de Venn da Figura 1.

Figura 1: Diagrama de Venn da *text mining* e suas interações com outros campos de estudo.



Fonte: Adaptado pelo autor (TALIB, et al., 2016).

Podemos notar que a Estatística e a Mineração de Texto estão diretamente ligadas com todas as áreas de estudo presentes na Figura 1. Vemos também que as duas áreas citadas juntamente ao Processamento de Linguagem Natural possuem dependência de recursos computacionais para seu desenvolvimento. Tendo como objetivo a aplicação de técnicas de Mineração de Texto, cada uma das áreas compõe partes do estudo apresentado neste trabalho de conclusão de curso.

2.1 História da Mineração de Texto

O entendimento e estudo da *text mining* teve início na metade dos anos 80, mas como não havia recursos computacionais e a análise demandava muito trabalho, muitas vezes com diversas falhas por conta da análise humana sujeita a erros, as técnicas e compreensão desses estudos só tiveram um grande salto por volta dos anos dois mil, avançando principalmente depois dos ataques terroristas de 11 de Setembro de 2001, quando a *data mining* e consequentemente a TM ganharam grande foco na luta contra o terrorismo (MIGUEL, 2019).

Outro ponto chave para história e evolução da TM, também nos anos dois mil, com o crescimento da IE (Inteligência Empresarial), houve um grande investimento no desenvolvimento de tecnologias computacionais e softwares, considerando que as empresas estavam aprendendo a utilizar dados para maximizar seus lucros (HEARST, 1999).

2.2 Processos de Mineração de Textos

Sabendo que os textos são dados não estruturados, um conjunto de técnicas foi desenvolvido para analisá-los em sua complexidade de forma que os resultados sejam verdadeiros. Esse conjunto de ferramentas e técnicas pertencem a área de Descoberta de conhecimentos em Textos (*Knowledge Discovery frow Text - KDT*) (PALAZZO, et al., 2006). Atualmente a área tem grande repercussão na análise qualitativa e quantitativa para grandes volumes de texto que podem ser extraídos de e-mails, páginas da Web, arquivos em diversos formatos, campos textuais em bases de dados e textos eletrônicos. Os processos da TM estão presentes no fluxograma da Figura 2.

Figura 2: Fluxograma das etapas de Mineração de Texto.

Fonte: Morais e Ambrósio (2007).

O processo como mostra a Figura 2 tem início na seleção dos textos, ou seja, na escolha dos dados que serão trabalhados, a partir daí é possível seguir para dois tipos de análise, sendo a semântica, com finalidade de desenvolvimento de tecnologias de interface humana, ou a análise estatística com o objetivo de extração de informações. Conhecendo seu objetivo, os dados são tratados de acordo com o tipo de análise escolhida, passando pela indexação e normalização onde tratamos os dados para que fiquem com uma característica única nos caracteres, a partir daí chamamos essa estrutura de texto normalizada de *corpus*.

Com o *corpus* pronto, a análise de relevância ou frequência dos termos pode ser realizada utilizando técnicas estatísticas ou semânticas. Por fim, com os termos mais relevantes é possível aplicar outras técnicas que atendem aos objetivos do estudo.

2.2.1 Análise Semântica

A análise semântica está diretamente associada ao Processamento de Linguagem Natural (PLN) que a avaliam as sequências dos termos em um texto com

a função de entender seu propósito (ROSA, 1998). Uma série de conhecimentos são necessários para o PLN ser válido, sendo eles:

- Conhecimento Morfológico: É o entendimento da estrutura do idioma.
- Conhecimento Sintático: É o conhecimento das palavras e como juntas podem produzir sentenças.
- Conhecimento Semântico: É o entendimento da palavra e seu significado.
- Conhecimento Pragmático: É o uso do idioma em diferentes sentidos e contextos e como afetam a interpretação.
- Conhecimento do Discurso: É o entendimento de como as sentenças anteriores podem determinar as próximas.
- Conhecimento do Mundo: É o conhecimento geral do domínio do idioma e suas variações.

A análise Semântica está diretamente ligada com as Inteligências Artificiais (IA), sejam elas assistentes virtuais inteligentes que respondem a comando de voz, texto preditivo no teclados de digitação, chamadas telefônicas digitais ou mesmo em filtros do e-mail que classificam possíveis spams. Contudo a análise semântica e o PLN tem o intuito de colaborar com a aprendizagem de máquinas de uma forma geral.

2.2.2 Análise Estatística

A análise estatística para a mineração de texto se baseia em não considerar o texto como um todo e sim na frequência com que as palavras aparecem nos documentos, ou seja, o texto é dividido em palavras únicas desconsiderando pontuações, acentos e palavras de conjunção e são contabilizadas as palavras relevantes ao texto e sua frequência, possibilitando a visualização gráfica das palavras chaves dentre outras análises. Suas etapas são:

- Codificação dos dados: Nesse processo os dados são tratados dividindo o texto em palavras e removendo tudo aquilo que é irrelevante para a análise. Entretanto é necessário ter cuidado, pois pode haver perda de informações que são relevantes e, com isso, é importante conhecer e entender o problema para evitar erros. Pequenas análises também são realizadas nessa etapa

como a verificação das palavras mais relevantes e se estão escritas da forma correta, para não serem contabilizadas duas vezes.

- Estimativa dos dados: Nessa etapa é importante entender os dados e traçar os objetivos para a escolha e definição de um modelo que se aplique aos dados e seja capaz de corresponder aos objetivos.

- Modelos de Representação de Documentos: Existem diversos modelos para a análise de documentos que são capazes de associar palavras a documentos e técnicas que associam a relevância da palavra para o documento. Ainda existem falhas nesses modelos por serem recentes quando comparados a modelos Estatísticos clássicos que trabalham com dados numéricos como modelos lineares ou de regressão, contudo esses modelos vem se aprimorando de acordo com o crescimento e a necessidade do estudo de texto.

3 Organização dos Dados

Para a análise e comparação dos resultados quatro pré-candidatos a presidência do Brasil foram escolhidos, identificados pelas letras A, B, C e D, sem a divulgação dos nomes para que não exista tendências ou conclusões prévias. Os dados de suas contas do *Twitter* serão analisados como um todo, desde as postagens dos pré-candidatos até as respostas dos seguidores e não seguidores da conta.

Tendo como objetivo utilizar dados do *Twitter* para realizar as análises, algumas etapas são necessárias para o processo como a escolha do software e pacotes que serão utilizados para as análises, desenvolvimento de uma ferramenta de coleta dos dados nesse *software*, uma conta de desenvolvedor da rede e suas chaves de acessos. Além disso, é fundamental que os dados sejam organizados e filtrados da maneira correta.

3.1 *Softwares* e Pacotes

Utilizaremos o *software* Rstudio para a coleta e realização das análises estatísticas e o Excel para armazenamento desses dados.

3.1.1 Excel

O Excel ou Microsoft Excel é um aplicativo de criação de planilhas eletrônicas e foi desenvolvido pela Microsoft em 1987. O aplicativo consiste em linhas e colunas que permitem a criação, manipulação e apresentação de bases de dados, de acordo com a necessidade do usuário.

Dado nosso objetivo a planilha será utilizada para armazenar dados coletados por meio do *software* R e a versão utilizada será Excel 14.0.

3.1.2 Rstudio

O Rstudio é uma versão iterativa do *software* R que foi desenvolvido em 1996, pelos professores de estatística Ross Ihaka e Robert Gentleman, da Universidade de Auckland. O *software* tem uma linguagem de programação similar a linguagem S desenvolvida por John Chambers, seu número de usuários vem crescendo exponencialmente devido ao fato de ser um software gratuito com uma linguagem acessível, estar disponível para as plataformas Windows, Linux e Mac e conta com mais de 15 mil pacotes que auxiliam nas análises. Utilizaremos a versão 8.16 build 180499 2021.

Os pacotes que utilizaremos para realizar a análise são:

- **rtwitte:** O pacote possibilita a coleta de dados por diversas funções, sendo elas, coletar dados de um perfil específico, uma determinada palavra ou até mesmo uma hashtag.
- **wordcloud:** Sua função principal é contabilizar a quantidade de vezes que a palavra aparece no banco de dados e realizar uma nuvem de palavras com os termos mais frequentes.
- **tidytext:** O pacote realiza algumas funções da mineração de texto, como separar as palavras e pode ser combinado com outros pacotes como o ggplot2 e dplyr para análises gráficas.
- **stringr:** O pacote também executa funções de mineração de texto, no caso usaremos especificamente para a limpeza do banco de dados.
- **tidyverse:** Tem a função de carregar outros pacotes, incluindo pacotes de filtros acessados diretamente da Web, alguns desses filtros ajudaram na limpeza dos dados.

- **tm:** Possuí funções de mineração de texto e é suporte para aplicativos e outros pacotes.
- **dplyr:** É responsável pela manipulação dos dados e auxilia na construção da base final, pertence ao grupo de pacotes do tidyverse.
- **xlsx:** Possibilita carregar a base de dados montada no R para o Excel sem alterar sua estrutura ou dados.

3.2 Ferramenta de Coleta dos Dados

Para conseguir acesso aos dados do *Twitter* foi necessário "transformar' uma conta normal em uma conta de desenvolvedor do app, para obter as chaves de acessos (APIs) necessárias. O primeiro passo é acessar o link developer.twitter.com, a partir daí é possível selecionar quais suas intenções ao adquirir as APIs, dentre as opções estão "*Student*", "*Doing Academic Research"*, "*Teaching*", entre outras que envolvem customizar sua própria conta.

O próximo passo para adquirir sua conta de desenvolvedor é responder algumas perguntas, como, seu nome, país, e-mail, etc. Por fim o *Twitter* pede que você escreva exatamente o que pretende realizar com as APIs e os tipos de análise. Se for para fins acadêmicos é necessário citar o nome da instituição e curso matriculado. Após esses passos o Twitter leva cerca de dois dias para aprovar sua solicitação e fornecer a conta de desenvolvedor.

Tendo acesso a sua conta de desenvolvedor, passamos para o *Software* Rstudio para a criação da ferramenta de coleta. Por meio do pacote "rtwitte" cadastramos e validamos nossas APIs, sendo elas: consumer_key, consumer_secret, access_token e access_token_secret.

Com as chaves cadastradas podemos realizar a busca por dados usando funções do pacote "rtwitte".

3.2.1 Coleta dos Dados

Os dados dos quatro candidatos serão coletados do dia 8 de Setembro de 2021 ao dia 27 de Outubro de 2021 em intervalos de uma semana, cujos dias estão destacados no Quadro 1.

Quadro 1: Data para a coleta de dados dos quatro pré-candidatos.

Dias em que os Dados serão coletados				
Setembro	08/09/2021	15/09/2021	22/09/2021	29/09/2021
Outubro	06/10/2021	13/10/2021	20/10/2021	27/10/2021

Fonte: Elaborado pelo autor.

Em cada dia de coleta serão recolhidos cerca de 10 mil tweets de cada pré-candidato e armazenados em um banco de dados no Excel. Caso o candidato não tenha atingido 10 mil tweets em uma semana esse número será reduzido a quantidade necessária para que o mesmo tweet não esteja presente duas vezes no banco de dados alterando a análise. Ao fim das 8 semanas a análise dos dados e aplicação do modelo será realizada.

3.3 Tratamento dos Dados

Ao coletar os dados do *Twitter* as seguintes variáveis constam no primeiro banco de dados não tratados como indica o Quadro 2.

Quadro 2: Dicionário de Variáveis.

Dicionário de Variáveis	
Nome do Campo	Descrição
Text	Texto
favorited	Favoritou (referente a conta de desenvolvedor)
favoritedCount	Contagem de favoritos
replyToSN	Respondito a quem
Created	Criado no formato YYYY-MM-DD
replyToSID	Respondido a (código do Sistema de Identificação, 19 dígitos)
id	Identificação (19 dígitos)
replyToUID	Respondido a (Identificação do usuário, vária de 11 a 19 dígitos)
statusSource	Fonte do tweet (endereço URL)
screenName	Nome do usuário
retweetCount	Contagem de retweets
isRetweet	É um retweet (Verdadeiro ou falso)
retweeted	Retweetou (referente a conta de desenvolvedor)
longitude	Longitude
Latitude	Latitude

Fonte: Elaborado pelo autor.

Como nosso objetivo é a análise dos textos presentes em cada *tweet* vamos considerar somente a variável que contém o texto necessário (*text*). As variáveis descartadas para essa análise são muito úteis em outros tipos de estudo como, por

exemplo, o número de *retweets*. A latitude e longitude podem ser usadas por empresas para avaliar o alcance de certa propaganda de um produto ou divulgação de campanhas da marca.

Os textos que utilizaremos nas análises são considerados informais, pois estão presentes em uma rede social de caráter pessoal. Logo para a análise estatística da TM (Figura 2), precisamos realizar a normalização dos caracteres e descartar o que for irrelevante. Para exemplificar, vamos supor o seguinte *Tweet* no banco de dados:

RT @joao_souza: @ana_silva O dia hoje está Lindo!!! Vamos tomar um Sorvete? #Sorve(teria) http://sorveteria.com.br

A ferramenta de coleta estrutura o *tweet* recolhido da forma indicada pela Figura 3.

Figura 3: Estrutura da variável "*text*" no banco de dados.

Fonte: Elaborado pelo autor.

Como trabalhamos com textos informais a mesma palavra pode ser considerada diferente se um usuário escreveu com letra maiúscula e o outro não. Portanto, utilizando funções do pacote "tidyverse" primeiramente convertemos o texto inteiro para letras minúsculas.

rt @joao_souza: @ana_silva o dia hoje está lindo!!! vamos tomar um sorvete? #sorve(teria) http://sorveteria.com.br

O próximo passo é remover os usuários contidos no texto.

rt : o dia hoje está lindo!!! vamos tomar um sorvete? #sorve(teria) http://sorveteria.com.br

Endereços de sites não são relevantes portanto podemos retirar.

rt : o dia hoje está lindo!!! vamos tomar um sorvete? #sorve(teria)

Retiramos então as pontuações, levando em conta que queremos a análise estatística que considera a frequência das palavras. No caso de uma análise semântica as pontuações seriam necessárias.

rt o dia hoje está lindo vamos tomar um sorvete #sorve(teria)

Por fim limpamos caracteres especiais presentes no texto.

rt o dia hoje está lindo vamos tomar um sorvete sorveteria

3.3.1 *Stopwords*

Segundo Berry e Kogan (2010) "as *stopwords* são consideradas não informativas, uma vez que não resumem o conteúdo abordado pelo texto de forma satisfatória". Consideramos então as *stopwords* palavras que podem ser excluídas dos textos na TM, pois não trazem informações relevantes a análise. Essas palavras geralmente são artigos, conjunções, preposições, pronomes e alguns verbos comumente utilizados. Entretanto outras palavras podem ser incluídas dado que cada texto possuí um assunto.

O pacote “stringr” possui uma lista de *stopwords* da língua portuguesa, algumas das palavras estão contidas no Quadro 3.

Quadro 3: Stopwords.

Rt	sua	tinha	meus	estivera	houverão	tenho
de	ou	foram	Minhas	estivéramos	houveria	tem
a	ser	essa	teu	esteja	houveríamos	temos
o	quando	num	Tua	estejamos	houveriam	tém
que	muito	nem	Teus	estejam	sou	tinha
e	há	suas	Tuas	estivesse	somos	tínhamos
do	nos	meu	nosso	estivéssemos	são	tinham
da	já	às	Nossa	estivessem	era	tive
em	está	minha	Nossos	estiver	éramos	teve
um	eu	têm	Nossas	estivermos	eram	tivemos
para	também	numa	dela	estiverem	fui	tiveram
é	só	pelos	delas	hei	foi	tivera
com	pelo	elas	esta	há	fomos	tivéramos
não	pela	havia	estes	havemos	foram	tenha
uma	até	seja	estas	hão	fora	tenhamos
os	isso	qual	aquele	houve	fôramos	tenham
no	ela	será	aquela	houvemos	seja	tivesse
se	entre	nós	aqueles	houveram	sejamos	tivéssemos
na	era	tenho	aquelas	houvera	sejam	tivessem
por	depois	lhe	isto	houvéramos	fosse	tiver
mais	sem	deles	aquilo	haja	fôssemos	tivermos
as	mesmo	essas	Estou	hajamos	fossem	tiverem
dos	aos	esses	Está	hajam	for	terei
como	ter	pelas	Estamos	houvesse	formos	terá
mas	seus	este	Estão	houvéssemos	forem	teremos
foi	quem	fosse	Estive	houvessem	serei	terão
ao	nas	dele	Esteve	houver	será	teria
ele	me	tu	estivemos	houvermos	seremos	teríamos
das	esse	te	estiveram	houverem	serão	teriam
tem	eles	vocês	Estava	houverei	seria	pra
à	estão	vos	estávamos	houverá	seríamos	para
seu	você	lhes	Estavam	houveremos	seriam	vamos

Fonte: Elaborado pelo autor.

Utilizando o mesmo exemplo da Seção 3.3, após a remoção dos caracteres, faremos a retirada das *stopwords*. Ficamos então com o seguinte corpo de texto que são as palavras relevantes para a análise:

dia hoje lindo tomar sorvete sorveteria

4 Técnicas para Análise dos Dados

Com o banco de dados tratado, podemos dar início às análises estatísticas considerando principalmente a frequência dos termos nos *tweets* de cada conta dos pré-candidatos A, B, C e D.

Vamos definir como documento a conta dos quatro pré-candidatos, ou seja, o documento A se refere ao banco de dados do pré-candidato A, o documento B é referente ao banco de dados do pré-candidato B, e assim sucessivamente.

Uma análise descritiva simples será aplicada a cada documento contemplando uma visualização gráfica dos termos mais frequentes, a fim de verificar se alguma outra palavra deve ser incluída como stopword, ou se devemos entender o contexto e sua relevância para o documento e mantê-la no banco de dados.

Após estudar e entender o contexto de cada documento aplicaremos outras técnicas como uma *word cloud* que possibilita uma visualização interativa dos termos mais frequentes, uma análise dos sentimentos, a técnica TF-IDF para identificar as palavras mais importantes em cada documento, um dendrograma e por fim a aplicação do modelo *Latent Dirichlet Alocation*.

4.1 Nuvem de Palavras (*Word Cloud*)

A nuvem emergiu a partir da análise lexical, entendendo como léxico o conjunto de palavras que compõe um determinado texto. Por esse ângulo, a técnica de construção destas nuvens consistiu em usar tamanhos e fontes de letras diferentes de acordo com a frequência das ocorrências das palavras no texto analisado (RIVEDENEIRA, et al., 2007).

A nuvem de palavras geralmente é usada como recurso visual, quando o objetivo é destacar as palavras que ocorrem com maior frequência nos documentos. O cálculo das frequências dos termos (TF) é realizado da seguinte forma:

$$\text{TF(termo)} = \left(\frac{Número\ de\ vezes\ que\ o\ termo\ aparece\ no\ documento}{Número\ total\ de\ termos\ do\ documento}\right) \quad (1)$$

Utilizando funções do pacote "wordcloud" podemos mensurar a nuvem escolhendo a frequência mínima e máxima dos termos que estarão presentes na nuvem.

4.2 Análise de Sentimentos

Liu em 2012 definiu a análise de sentimentos como “Área de estudo que analisa as opiniões, sentimentos, avaliações, apreciações, atitudes e emoções das pessoas em relação a entidades como produtos, serviços, organizações, indivíduos, questões, eventos, tópicos e todos os seus atributos relacionados” (LIU, 2012).

Dentro do contexto semântico a análise de sentimentos define as expressões de sentimentos, polaridade da expressão e sua relação com o assunto. Por exemplo, na frase "Pedro bateu em Carlos", o verbo "bateu" denota um sentimento positivo em relação a Pedro e um sentimento negativo em relação a Carlos (exemplo adaptado de NASUKAWA, et al., 2003).

Podemos então dizer que análise de sentimentos dentro do contexto da mineração de texto estatística tem por objetivo identificar palavras que já possuam um sentimento atribuído como, por exemplo, a palavra “feliz”, que denota um sentimento positivo ou a palavra “triste”, que está associada a um sentimento negativo. Com o avanço da aprendizagem de máquinas já existem pacotes capazes de identificar diversos sentimentos como medo, confiança, alegria, raiva, entre outros.

4.3 Análise de Agrupamentos

A análise de agrupamentos envolve técnicas estatísticas multivariadas para agrupar dados da melhor maneira. Dessa forma, dada uma amostra de n objetos, cada um deles medido por p variáveis, procura-se um esquema de classificação que agrupe os objetos em k grupos (BUSSAB,1990).

4.3.1 Métodos Hierárquicos Aglomerativos

Os métodos Hierárquicos podem ser subdivididos em aglomerativos e de partição. De acordo com Anderberg (1973), existem quatro etapas a se seguir.

1- Inicia-se com N grupos cada um contendo um elemento;
2- Determinar a Matriz de dissimilaridade e o par de grupos com menor dissimilaridade;
3- Os dois grupos se unem e se calcula novas medidas até que o número total de grupos é reduzido a um;

4- Executam os passos 2 e 3 (N-1) vezes, até que todos os elementos estejam alocados.

4.3.1.1 Método da Ligação Completa (*Complete Linkage - CLINK*)

Definimos o método como a maior distância entre um elemento de G_1 (Grupo 1) e um elemento de G_2 (Grupo 2) Silvestre (2006 *apud* Silvestre, 2016).

$$d[G_1,G_2]=\max d_{ik} \quad \text{para } i \in G_1 \text{ e } k \in G_2 \tag{2}$$

Utilizando o método da ligação completa será elaborado um dendrograma dos dados dos quatro documentos. O dendrograma nos ajudará posteriormente a identificar as quantidades de assunto (tópicos) em cada documento.

4.4 *Term Frequency – Inverse Document Frequency* (TF-IDF)

Parte importante da TM e do PLN é poder quantificar do que se trata o texto. Uma possibilidade para isso é o calculo da frequência dos termos (TF) apresentado em 4.1, onde consideramos a remoção das *stopwords* para facilitar a identificação dos termos chaves.

Uma abordagem onde as *stopwords* não são retiradas é a Frequência Inversa do Documento de um termo (IDF), que diminui o peso das palavras normalmente usadas e aumenta o peso das palavras que não são muito usadas em um grupo de documentos. (SILGE e ROBINSON, 2017). A frequência inversa do documento é obtida por meio do cálculo:

$$\text{IDF(termo)} = \ln\left(\frac{Número\ de\ documentos}{Número\ de\ documentos\ que\ contém\ o\ termo}\right) \tag{3}$$

Através da multiplicação das medidas TF e IDF obtêm-se a estatística TF-IDF que mede a importância de um termo para o documento dentre um grupo de documentos, ou seja, conseguimos destacar as palavras que são realmente relevantes em cada um dos documentos (AIZAWA, 2002).

Por exemplo, ao pegar uma coleção de livros de romance escritos pelo mesmo autor e aplicar a estatística TF-IDF, os termos relevantes para cada livro serão os que apresentarem maior TF-IDF, ou seja, os nomes dos personagens, pois

são as palavras que não estão presentes em outros documentos e são significativas para cada livro (Exemplo adaptado de SILGE e ROBINSON, 2016).

Nosso objetivo ao aplicar a técnica é descobrir os termos que diferenciam cada pré-candidato, dado o contexto político presente nos documentos.

5 Modelos Probabilísticos de Tópicos

Com a necessidade do extrair informações importantes de textos, uma nova área surgiu em 2003, chamada de modelos probabilístico de tópicos (Probabilistic Topic Models). Sua intenção é buscar relações latentes entre documentos e termos que sejam importantes na descoberta de informações (BLEI, 2011).

Griffiths e Steyvers em 2004 definiram os modelos probabilísticos de tópicos como "Um conjunto de algoritmos cujo objetivo é descobrir estruturas temáticas ocultas em grandes grupos de documentos" (GRIFFITHS e STEYVERS, 2004). Esses modelos foram propostos inicialmente para análises de texto, mas passaram a ser usados em análises de imagens, grafos, dentre outras áreas (SIVIC, et al., 2005).

A área de estudo se fortaleceu também em 2003, quando Blei, Ng e Jordan apresentaram o modelo base *Latent Dirichlet Allocation* (LDA).

5.1 *Latent Dirichlet Allocation* (LDA)

O *Latent Dirichlet Allocation* é um modelo generativo para grupos de dados discretos como corpus de texto (BLEI; NG; JORDAN, 2003). Modelos generativos são aqueles que aleatoriamente geram os dados a partir das variáveis latentes. O LDA é um modelo probabilísticos onde os dados observáveis são os termos de cada documento e os não observáveis são os tópicos presentes em cada documento, os seus hiper-parâmetros são dados *a priori* no modelo.

O LDA considera dois princípios na sua fundamentação, sendo cada documento uma mistura de tópicos, ou seja, cada documento possui uma mistura de assuntos/tópicos. Esses tópicos serão identificados através do dendrograma explicado na Seção 4. O outro princípio base é que cada tópico é uma mistura de palavras que são os termos analisados anteriormente. O modelo usa desses princípios para estimar ao mesmo tempo uma mistura de palavras que está associada a cada tópico e determina uma mistura de tópicos que descreve cada documento, ambos são distribuídos obedecendo a distribuição de Dirichlet.

Temos a função de densidade da distribuição de Dirichlet, denotada como $Dir(z, \alpha)$:

$$Dir(z, \alpha) = \frac{1}{B(\alpha)} \prod_{k=1}^{K} z_k^{\alpha_k - 1}, \tag{4}$$

onde $z = (z_1, \dots, z_k)$ é uma variável K-dimensional com $0 \leq z \leq 1$ e $\sum_{i=1}^{K} z_i = 1$. O hiper-parâmetro da distribuição é $\alpha = (\alpha_1, \dots, z_K)$ e a função $B(\alpha)$ é a função Beta expressa em termos da função gama Γ:

$$B(\alpha) = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{K} \alpha_K)} \tag{5}$$

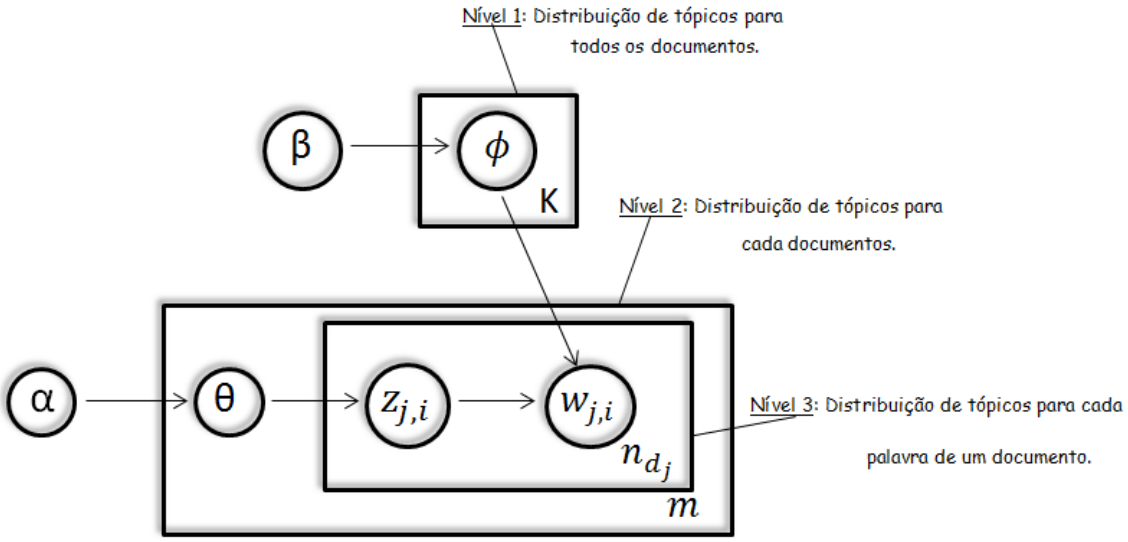
A distribuição Dirichlet é a priori conjugada da distribuição multinomial, por isso é muito utilizada na inferência Bayesiana (BISHOP, 2006).

Para entender o processo gerador do modelo LDA, vamos considerar d_j como um documento criado com distribuições $\phi_k \sim Dir(\phi_k, \beta)$ com $0 \leq k \leq K$ e $\theta_j \sim Dir(\theta, \alpha)$, onde ϕ é uma variável n-dimensional sendo n o número de palavras do vocabulário e θ uma variável K-dimensional sendo K o número de tópicos. ϕ e θ são distribuições de probalibilidades então $\sum_j^n \phi_j = 1, \phi > 0$, e $\sum_i^k \theta_i = 1, \theta > 0$. Ambas as variáveis são geradas da distribuição Dirichlet com seus hiper-parâmetros β e α.

O LDA é um modelo hierárquico de três níveis, onde no primeiro nível temos a distribuição de tópicos para todos os documentos, no nível dois passamos pra distribuição de tópicos para cada um dos documentos e no nível três temos a distribuição de tópicos para cada termo de cada documento. Uma característica do modelo é que cada tópico possui sua própria distribuição θ_j.

Podemos visualizar o processo generativo do modelo LDA apresentado na Figura 4.

Figura 4: Fluxograma do Modelo *Latent Dirichlet Allocation*.



Fonte: Elaborado pelo autor (Adaptado de Blei; Ng; Jordan, 2003).

As notações do modelo são definidas como:

- **β** - Parâmetro da probabilidade a priori de Dirichlet nas distribuições de termos por tópicos;
- **α** - Parâmetro da probabilidade a priori de Dirichlet nas distribuições de tópicos por documento;
- $\boldsymbol{\phi}$ - Distribuição de palavras para o tópico k;
- **K** - Número de tópicos;
- **θ** - Distribuição de tópicos para o documento d_j;
- $\boldsymbol{z_{j,i}}$ - Distribuição de tópicos associado a palavra $w_{i,j}$ no documento d_j onde $1 \leq j \leq m$ e $1 \leq i \leq n$;
- $\boldsymbol{w_{j,i}}$ - Palavra observada no documento d_j, onde $1 \leq j \leq m$ e $1 \leq i \leq n$;
- $\boldsymbol{n_{d_j}}$ - Número de palavras do vocabulário;
- $\boldsymbol{m}$ - Número de Documentos.

Pelo fluxograma da Figura 4, notamos que os hiper-parâmetros α e β influenciam em todos os níveis do modelo. Intuitivamente interpretamos que um valor alto para α nos diz que existe uma maior mistura de tópicos por documento e um valor baixo que esses tópicos estão misturados entre os documentos, da mesma forma um valor alto para β indica que o tópico possui uma mistura de muitos termos, enquanto um valor baixo indica que o tópico é formado por poucos termos.

Notamos uma dependência grande entre as variáveis do modelo no processo generativo com isso é possível descrever a probabilidade de todas as variáveis latentes considerando as informações *a priori* (BLEI, 2011). Logo temos a seguinte distribuição de probabilidade conjunta:

$$p(z, w, \phi, \theta|\alpha, \beta) = \prod_{k=1}^{K} p(\phi_k|\beta) \prod_{j=1}^{M} p(\theta_j|\alpha) \left(\prod_{i=1}^{V} p(z_{j,i}|\vec{\theta}_j)\, p\left(w_{i,j}\middle|z_{i,j}, \phi_{z_{i,j}}\right) \right) \quad (6)$$

E pelo teorema de Bayes, formulamos a probabilidade de $p(z, w, \phi, \theta|\alpha, \beta)$ como o cálculo da *a posteriori* do modelo, temos então:

$$p(z, \phi, \theta|w, \alpha, \beta) = \frac{p(z, w, \phi, \theta|\alpha, \beta)}{p(w)} \quad (7)$$

Com o grande número de dependência, o problema computacional é inferir $p(z, w, \phi, \theta|\alpha, \beta)$, onde w são todas as palavras observadas no grupo de documentos. Uma possível resolução é inferir a probabilidade a posteriori de todo o modelo através da soma da distribuição conjunta de todos os valores (todos os termos da análise), entretanto o número de interações é exponencialmente grande (BLEI, 2011). Outro modo de resolver o problema é por meio de métodos de inferência para o modelo LDA, entre eles temos o método Amostrador de Gibbs (*Gibbs Sampling*) e método da Inferência Variacional (*Variational Inference*).

- **Amostrador de Gibbs:** O Amostrador de Gibbs é utilizado em diversos tipos de problemas na Física e Estatística, entre outras, ganhou popularidade pela facilidade de implementação (GEMAN; GEMAN, 1984).

Para a resolução do problema usaremos o Amostrador de Gibbs devido a sua facilidade de implementação como solução.

5.1.1 Amostrador de Gibbs

Levando em conta o resumo dado na seção **5.1** sobre o Amostrador de Gibbs, vamos entender como o método funciona. Vamos supor a variável K-dimensional não observada, $z = \{z_1, \dots, z_k\}$,onde z_i é o valor da i-ésima dimensão do vetor z e $z_{-i} = \{z_1, z_2, \dots, z_{i-1}, z_{i+1}, \dots, z_k\}$, queremos inferir $p(z|w)$, considerando w uma variável observada. O Amostrador de Gibbs faz a amostragem de cada dimensão condicionada a outras dimensões, ou seja, para inferir $p(z|w)$ o método retira amostras para cada dimensão i de z, onde z_i depende das amostras de z_{-i}. Esse processo é realizado T vezes até que exista a estimativa de $p(z|w)$ (Geman; Geman, 1984). As demonstrações de convergência do algoritmo estão presentes no trabalho de Russel e Norvig (2003).

Para o Modelo LDA o Amostrador de Gibbs deve retirar amostras de três variáveis, sendo elas z, θ e ϕ. No método o LDA é desenvolvido até amostrar a variável z, partindo de z encontra-se os valores de θ e ϕ.

O método introduz as seguintes variáveis contadoras:

- $\boldsymbol{c_{j,i,k}}$: Número de vezes que um termo w_i é atribuído ao tópico k em d_j;
- $\boldsymbol{c_{j,*,k}}$: Número de termos no documento d_j atribuído ao tópico k;
- $\boldsymbol{c_{*,i,k}}$: Números de vezes que o termo w_i é atribuído ao tópico k em todos os documentos;
- $\boldsymbol{c_{*,*,k}}$: Número de termos atribuídos ao tópico k considerando todo o grupo de documentos.

Essas variáveis contadoras são introduzidas para manter as características do modelo no processo de integração. O processo de integração nos da à equação de amostragem para o modelo LDA via o Amostrador de Gibbs, temos então a seguinte equação:

$$p\left(z_{a,b} = k \middle| z_{-(a,b)}, w, \alpha, \beta\right) \propto \frac{\left(c_{a,*,z_{a,b}}^{-(a,b)} + \alpha_{z_{a,b}}\right) \times \left(c_{*,w_{a,b},z_{a,b}}^{-(a,b)} + \beta_{w_{a,b}}\right)}{c_{*,*,k}^{-(a,b)} + \sum_{l=1}^{n} \beta_l} \tag{8}$$

Partindo da equação de amostragem da variável z, conseguimos a equação de estimação das estatísticas θ,

$$\theta_{j,k} = \left(\frac{c_{j,*,k} + \alpha_k}{n_{d_j} + \alpha_k}\right) \tag{9}$$

e também de ϕ,

$$\phi_{k,j} = \left(\frac{c_{*,w_i,k} + \beta_k}{c_{*,*,k} + n\beta_k}\right) \tag{10}$$

Com as equações 8,9 e 10, conseguimos realizar o processo de retiradas de amostras que contém toda informação necessária para o cálculo da *a posteriori* $p(z, \phi, \theta | w, \alpha, \beta)$ de estimação pelo método do Amostrador de Gibbs para o modelo LDA. Conseguimos estão as estimações das probabilidades de um termo estar presente em um tópico e as probabilidades de um tópico em um documento.

6 Resultados

6.1 Nuvem de Palavras

Após recolher e tratar os dados, utilizando o pacote wordcloud, conseguimos observar a partir das nuvens de palavras, apresentadas na Figura 5, os termos citados com maior frequência para cada candidato, desconsiderando as *stopwords*.

Figura 5: Nuvens de Palavras.

Fonte: Elaborado pelo autor.

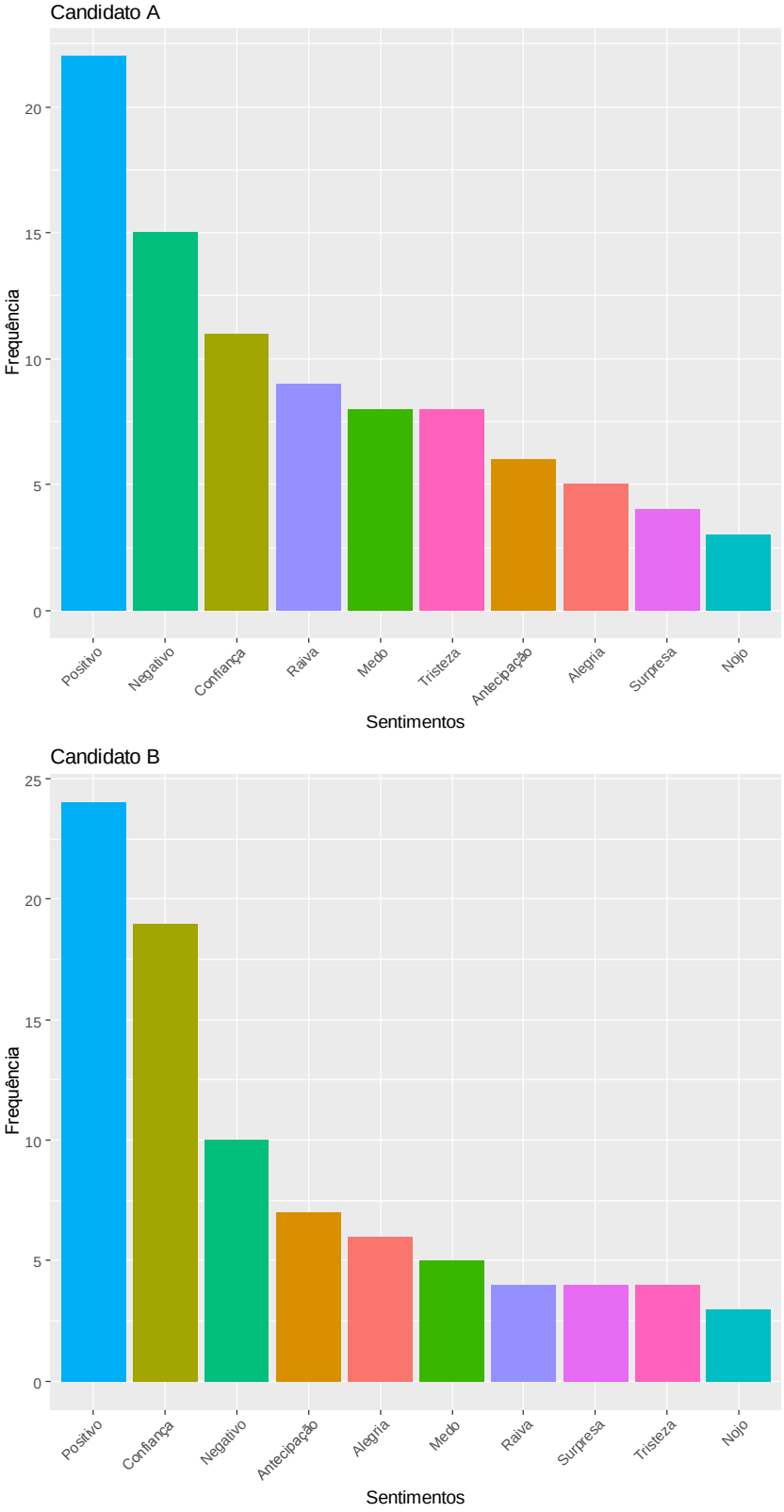
Notamos pela Figura 5 que os termos aposentados, democracia e constituição, são os mais citados no perfil do candidato A; para o candidato B os termos mais frequentes são Brasil, bom dia e Deus; já para o candidato C temos como o termo mais frequente o nome Bolsonaro e uma frequência menor de outros

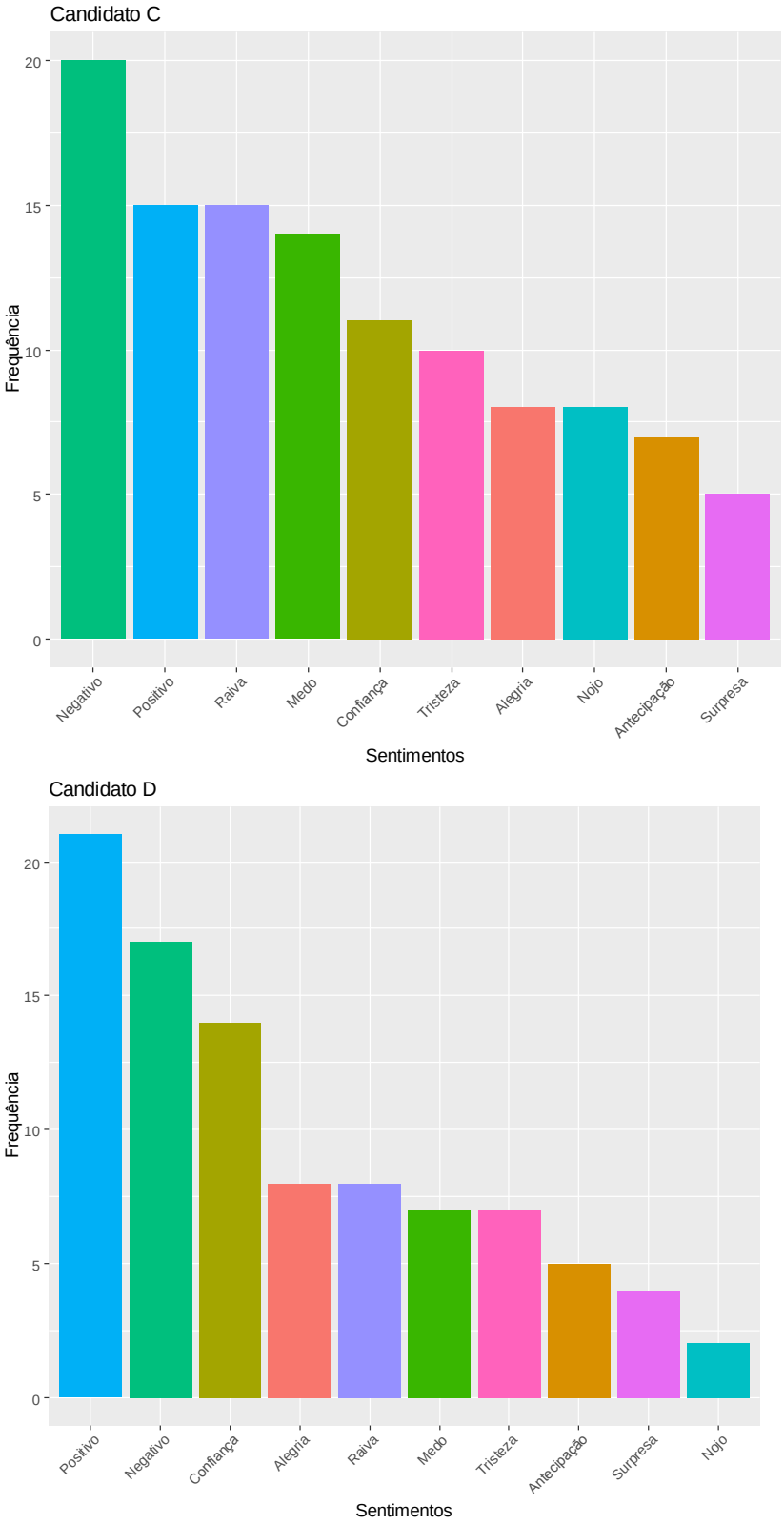
termos; entretanto para o candidato D notamos que não há um termo destacado como mais dito, porém existem vários termos com uma frequência alta quando comparado com os demais candidatos.

6.2 Análise de Sentimentos

Para realizar a análise de sentimento foi utilizado o pacote syuzhet que possui palavras associadas aos sentimentos, positivo, negativo, confiança, antecipação, medo, alegria, tristeza, surpresa, raiva e nojo. Considerando a base de dados sem as *stopwords*, obtivemos os seguintes resultados apresentados na Figura 5.

Figura 6: Análise de Sentimentos.





Fonte: Elaborado pelo autor.

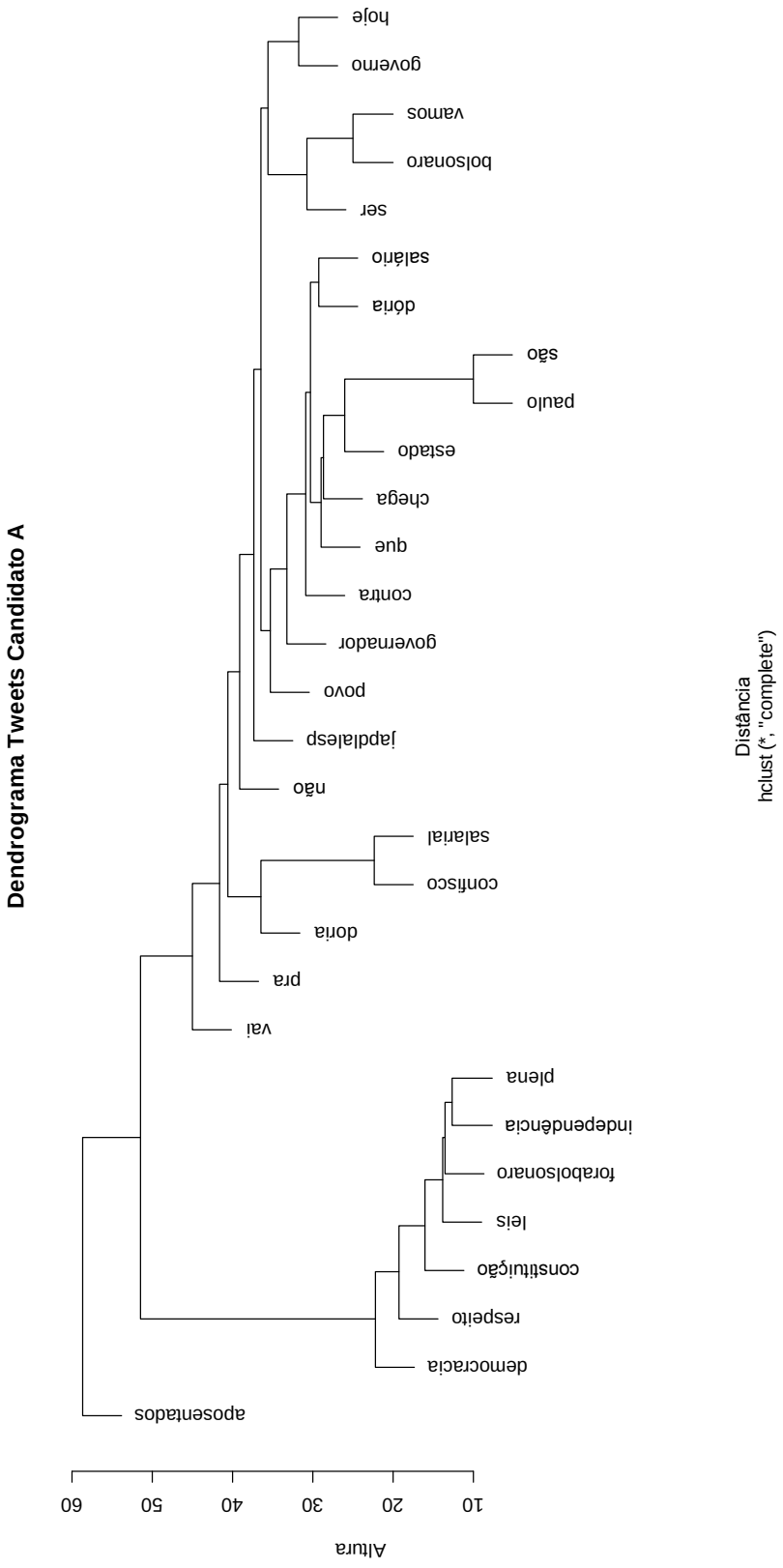
Podemos observar na Figura 6 que termos associados a sentimentos positivos são mais presentes nos candidatos A, B e D, enquanto os termos associados a sentimentos negativos estão mais associados ao candidato C. Entretanto, a análise

não é significativa ao considerarmos que poucas palavras foram associadas a algum sentimento, ou seja, o pacote utilizado ainda não reconhece uma quantidade considerável de palavras. Uma solução para isso é desenvolver uma escala quantitativa capaz de reconhecer uma quantidade maior de termos da língua portuguesa.

6.3 Método da Ligação Completa

O objetivo de aplicar o método da ligação completa era estimar através do dendrograma, conforme as Figuras 7,8,9 e 10 a quantidade de assuntos (tópicos) existentes dentro do perfil de cada candidato e verificar quais assuntos os perfis possuíam em comum e chegar a uma estimativa geral da quantidade de assuntos discutido no perfil dos 4 candidatos.

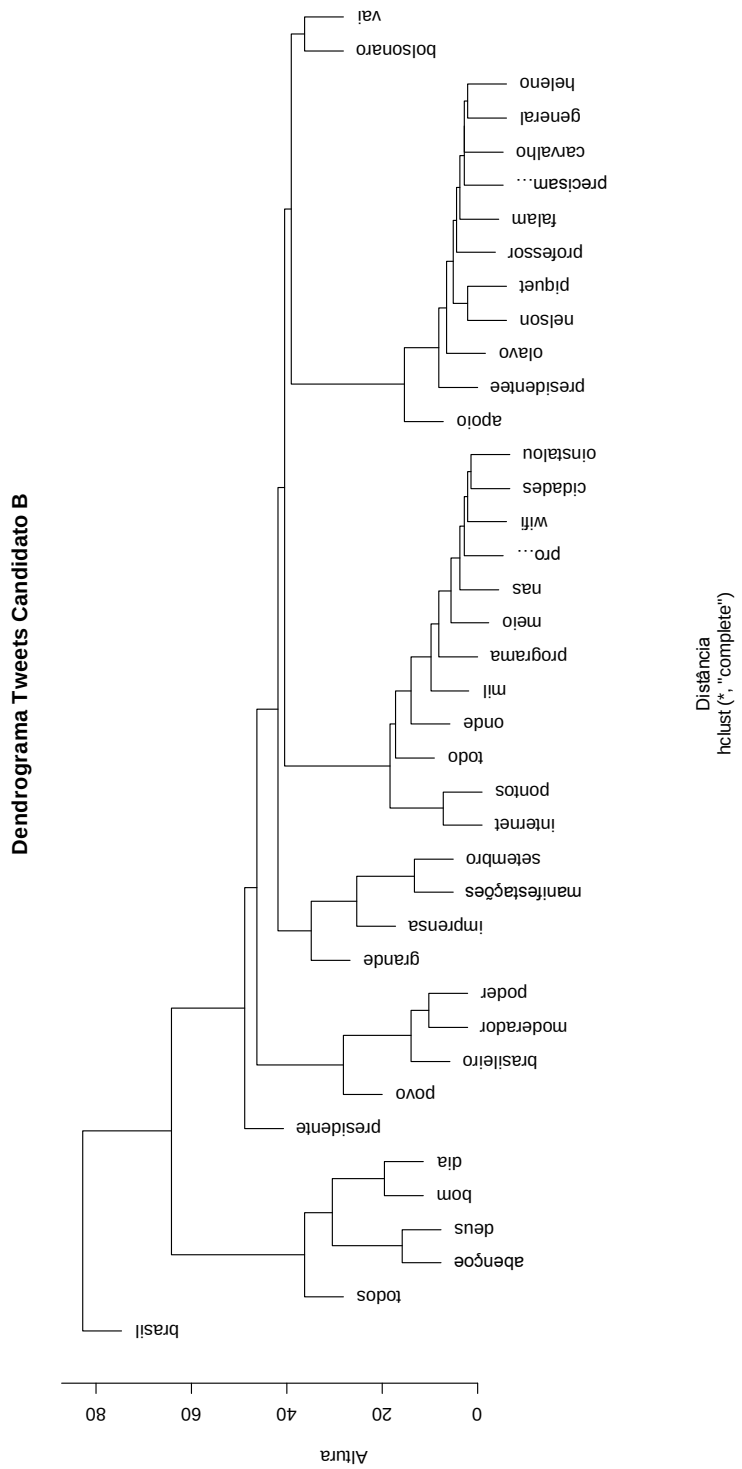
Figura 7: Dendrograma Candidato A.



Fonte: Elaborado pelo autor

Pelas ramificações do dendrograma do candidato A, ver Figura 7, identificamos aproximadamente quatro assuntos (tópicos) que estão relacionados ao governo e confisco salarial.

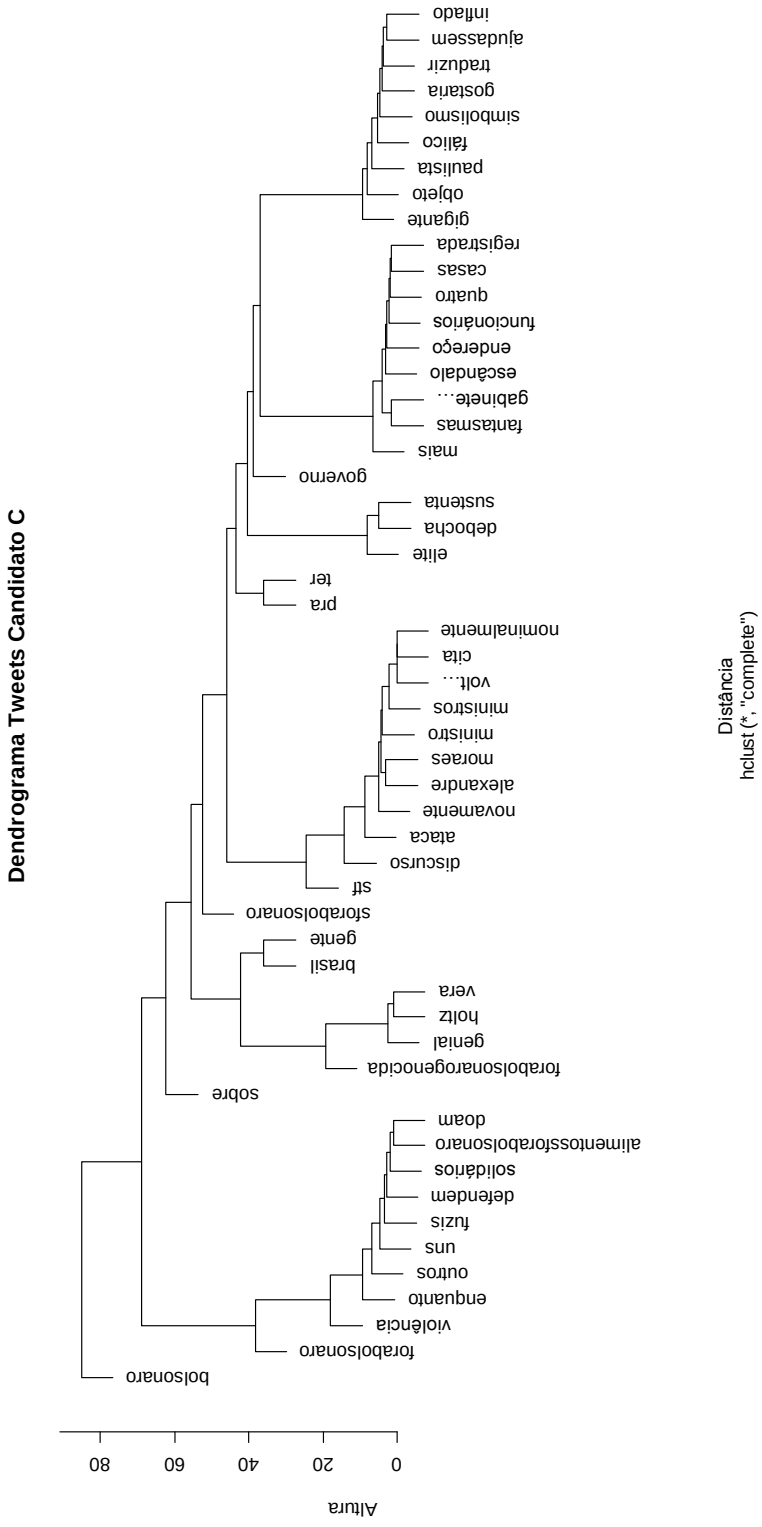
Figura 8: Dendrograma Candidato B.



Fonte: Elaborado pelo autor

No candidato B, ver Figura 8, temos um número maior de ramificações, entretanto todos os assuntos estão ligados de modo geral ao governo.

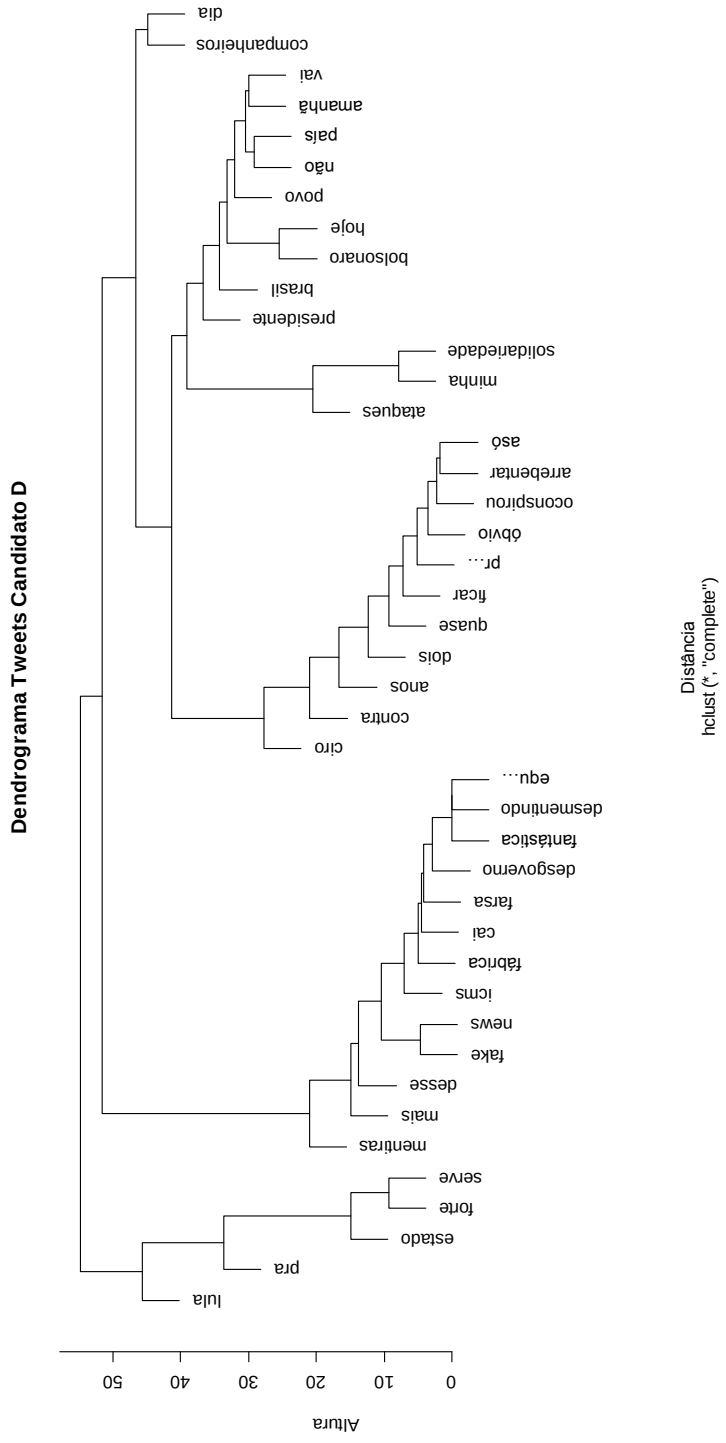
Figura 9: Dendrograma Candidato C.



Fonte: Elaborado pelo autor.

No dendrograma (Figura 9) do Candidato C, já identificamos aproximadamente 3 assuntos distintos, sendo o primeiro ligado a manifestações contra o governo atual, outro assunto ligado aos ministros e por fim uma discussão sobre as acusações e investigações sobre gabinetes fantasmas no país.

Figura 10: Dendrograma Candidato D.



Fonte: Elaborado pelo autor.

Por fim, no dendrograma apresentado na Figura 10 para o Candidato D, notamos uma ramificação mais expressiva em dois assuntos, um ligado diretamente a denúncias contra *fake news* e outro voltado a assuntos do governo.

Ao contabilizar todos os assuntos (tópicos) citados chegamos a uma estimativa de aproximadamente 8 tópicos discutidos com maior frequência entre os quatro candidatos. Então usaremos essa estimativa para o valor de k na aplicação do modelo LDA.

6.4 TF-IDF

Com o auxílio dos dendrogramas notamos que existem muitos termos em comum no perfil dos quatro candidatos por se tratarem de documentos com finalidade política. Uma forma de verificar os termos que são relevantes para cada candidato é aplicação da técnica TF-IDF, para isso vamos considerar a base de dados incluindo as *stopwords* para que a análise seja concisa, conforme ilustra a Figura 11.

Figura 11: Base de Dados Inicial.

Filter

	texto	Candidato
1	Eicom certeza você e o novo Aecio do PSDB o mauricin...	Candidato_A
2	...	Candidato_A
3	E pela metodologia dotambem	Candidato_A
4	É isso mesmo Não há Independência nem democracia ne...	Candidato_A
5	Vem pra rua João testa tua popularidade	Candidato_A
6	Quero verdeixarem as dive...	Candidato_A
7	Quem aumentou o ICMS de combustíveis alimentos e ...	Candidato_A
8	DeputadosPDLJa	Candidato_A
9	E pela metodologia dotambem	Candidato_A
10	Alô	Candidato_A
11	Lei q vc usa bem ne E gasta muito pra isso Mais de viat...	Candidato_A
12	Despoluir o rio e fácil quero ver eh conseguir recuperar ...	Candidato_A
13	DeputadosPDLJà Parece que no seu governo a lei e para...	Candidato_A
14	Comediante você ne	Candidato_A
15	Eicom certeza você e o novo Aecio do PSDB o mauricin...	Candidato_A

Showing 1 to 16 of 79,999 entries

Fonte: Elaborado pelo autor.

O primeiro passo para a aplicação do TF-IDF é quebrar os corpos de texto em termos separados e contabilizá-los por documento. Temos então o candidato, o termo, a quantidade de vezes que o termo aparece no perfil citado e o total de termos contidos naquele perfil, conforme ilustra a Figura 12.

Figura 12: Contagem de Termos.

```
#Bibliotecas de análise de texto
library(dplyr)
library(tidytext)

#Calculando a quantidade de palavras por documento
book_words <- Dados %>%                   #Esse bloco separa e conta as palavras por documento
  unnest_tokens(word, texto) %>%
  count(Candidato, word, sort = TRUE)

total_words <- book_words %>%             #Esse bloco agrupa os dados em uma tabela
  group_by(Candidato) %>%
  summarize(total = sum(n))

book_words <- left_join(book_words, total_words)
View(book_words)
```

Filter

	Candidato	word	n	total
1	Candidato_B	o	11349	232084
2	Candidato_C	e	10425	237963
3	Candidato_A	e	9663	217992
4	Candidato_C	de	9275	237963
5	Candidato_B	de	8683	232084
6	Candidato_B	a	7866	232084
7	Candidato_B	e	7609	232084
8	Candidato_D	de	7019	218227
9	Candidato_C	o	6861	237963
10	Candidato_A	o	6747	217992
11	Candidato_D	e	6747	218227
12	Candidato_D	que	6550	218227
13	Candidato_C	a	6375	237963
14	Candidato_D	o	6341	218227
15	Candidato_A	de	6309	217992

Showing 1 to 16 of 53,554 entries

Fonte: Elaborado pelo autor

Notamos que as *stopwords* são os termos mais frequentes em todos os documentos (perfis). Conhecendo o número de vezes que o termo é citado (n) e o total de termos de cada documento (total) podemos obter as estimativas TF, IDF e TF-IDF, descritas na Figura 13.

Figura 13: Estimativas TF, IDF e TF-IDF.

```
#------------------------------- TF-IDF ----------------------------------------#
book_tf_idf <- book_words %>%      #Calculando o TF, o IDF e o TF-IDF
  bind_tf_idf(word, Candidato, n)
View(book_tf_idf)
```

Filter

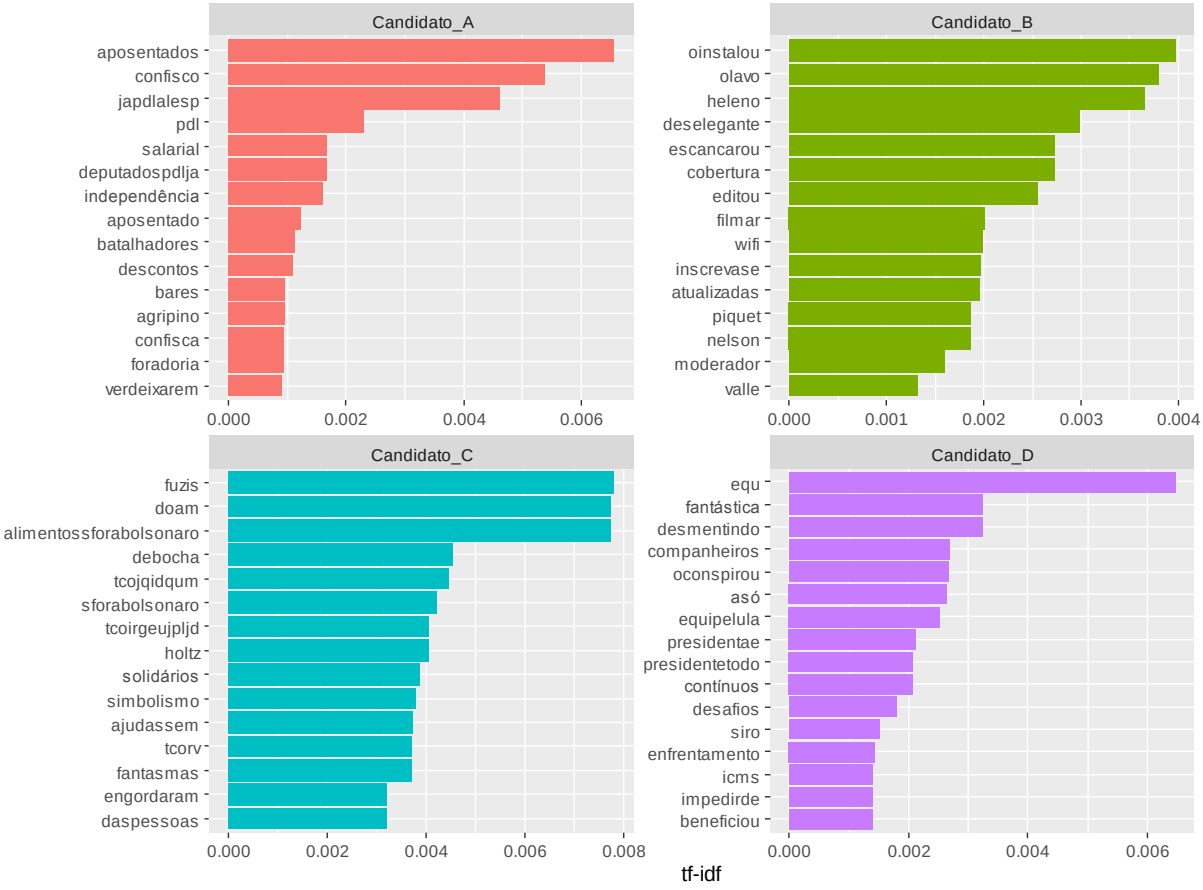
	Candidato	word	n	total	tf	idf	tf_idf
86	Candidato_C	fuzis	1339	237963	0.0056269252	1.3862944	0.0078005747
89	Candidato_C	doam	1330	237963	0.0055891042	1.3862944	0.0077481436
90	Candidato_C	alimentossforabolsonaro	1329	237963	0.0055849019	1.3862944	0.0077423179
46	Candidato_A	aposentados	2064	217992	0.0094682374	0.6931472	0.0065628820
129	Candidato_D	equ	1018	218227	0.0046648673	1.3862944	0.0064668793
155	Candidato_A	confisco	845	217992	0.0038762890	1.3862944	0.0053736776
194	Candidato_A	japdlalesp	724	217992	0.0033212228	1.3862944	0.0046041924
182	Candidato_C	debocha	778	237963	0.0032694158	1.3862944	0.0045323727
185	Candidato_C	tcojqidqum	764	237963	0.0032105832	1.3862944	0.0044508133
75	Candidato_C	sforabolsonaro	1451	237963	0.0060975866	0.6931472	0.0042265250
203	Candidato_C	holtz	695	237963	0.0029206221	1.3862944	0.0040488420
204	Candidato_C	tcoirgeujpljd	695	237963	0.0029206221	1.3862944	0.0040488420
219	Candidato_B	oinstalou	664	232084	0.0028610331	1.3862944	0.0039662340
88	Candidato_C	solidários	1333	237963	0.0056017112	0.6931472	0.0038828103
241	Candidato_B	olavo	637	232084	0.0027446959	1.3862944	0.0038049564

Showing 1 to 16 of 53,554 entries

Fonte: Elaborado pelo autor

Lembrando que os termos com maior TF-IDF são aqueles mais relevantes para cada candidato, podemos visualizar através dos gráficos de barras da Figura 14, os 15 termos mais relevantes para cada candidato considerando todos os candidatos.

Figura 14: Gráfico de barras dos 15 termos com maior TF-IDF de cada candidato.



Fonte: Elaborado pelo autor.

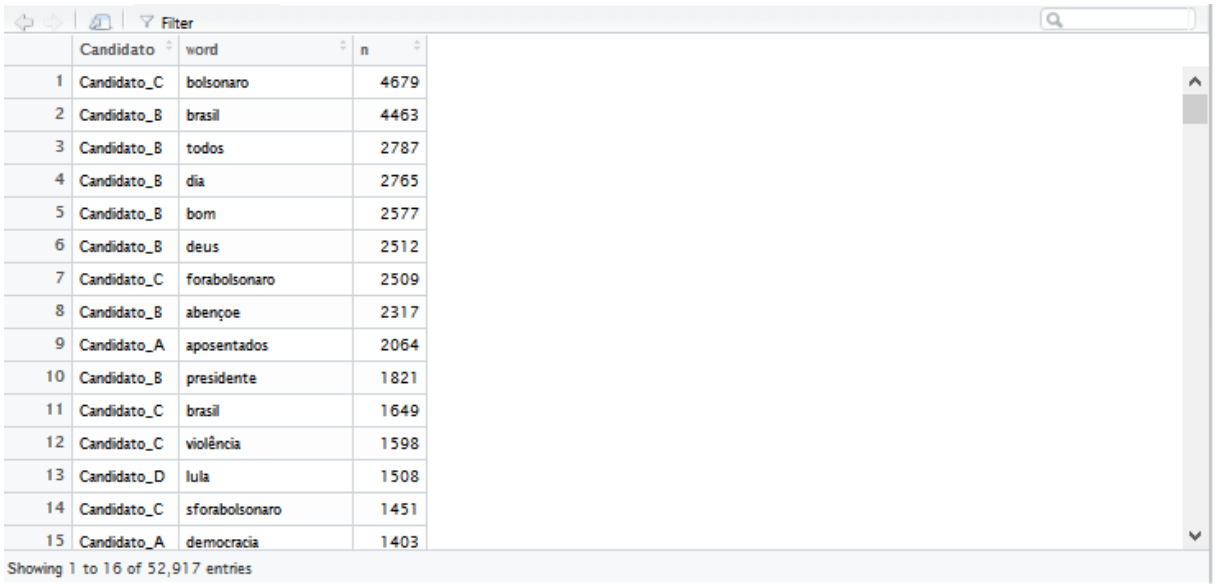
Através dos gráficos conseguimos verificar os termos que diferenciam os documentos tornando a técnica TF-IDF uma ferramenta fundamental para a tomada de decisões em outros contextos, como por exemplo, para uma marca de maquiagem que pretende contratar um influencer digital e precisa comparar e decidir dentre maquiadores que compartilham seu conteúdo, quais estão mais ligadas a identidade da empresa, ou com os objetivos de venda do produto.

6.5 Algoritmo de *Latent Dirichlet Allocation*

O LDA é um método popular para ajustar um modelo de tópico, pois considera cada documento como uma mistura de tópicos e cada tópico como uma mistura de termos. Na seção 6.4 já realizamos a contagem de termos em cada documento,

vamos utilizar essas informações na análise a seguir excluindo novamente as stopwords. Temos então a seguinte base de dados, apresentada na Figura 15.

Figura 15: Base de dados para aplicação do algoritmo LDA.



Filter

	Candidato	word	n
1	Candidato_C	bolsonaro	4679
2	Candidato_B	brasil	4463
3	Candidato_B	todos	2787
4	Candidato_B	dia	2765
5	Candidato_B	bom	2577
6	Candidato_B	deus	2512
7	Candidato_C	forabolsonaro	2509
8	Candidato_B	abençoe	2317
9	Candidato_A	aposentados	2064
10	Candidato_B	presidente	1821
11	Candidato_C	brasil	1649
12	Candidato_C	violência	1598
13	Candidato_D	lula	1508
14	Candidato_C	sforabolsonaro	1451
15	Candidato_A	democracia	1403

Showing 1 to 16 of 52,917 entries

Fonte: Elaborado pelo autor.

Primeiramente vamos transformar a base de dados em uma matriz de termos como apresentado no fluxograma do modelo (Figura 4). Isso permitirá estimar os parâmetros β e α que estão associados à probabilidade nas distribuições de termos por tópico e tópico por documento, respectivamente.

Usando a função LDA do pacote topicmodels vamos realizar as estimativas do parâmetro beta supondo o número de tópicos k=8 estimado na Seção 6.3 com o auxílio dos dendrogramas. A função LDA usa o método Amostrador de Gibbs para calcular o grande número de interações dependentes para a estimação de β e α. Temos então algumas estimativas de β, indicado na Figura 16.

Figura 16: Estimativas de β.

```
#----Distribuição de palavra por tópico-----#
#Transformando em uma Matriz de termos
desc_dtm <- word_counts %>%
  cast_dtm(Candidato, word, n)
desc_dtm

#Fazendo as estimações de beta
library(topicmodels)
desc_lda <- LDA(desc_dtm, k = 8, control = list(seed = 1234))
desc_lda
tidy_lda <- tidy(desc_lda)
tidy_lda

top_terms <- tidy_lda %>%
  group_by(topic) %>%
  slice_max(beta, n = 15, with_ties = FALSE) %>%
  ungroup() %>%
  arrange(topic, -beta)
view(top_terms)
```

Filter

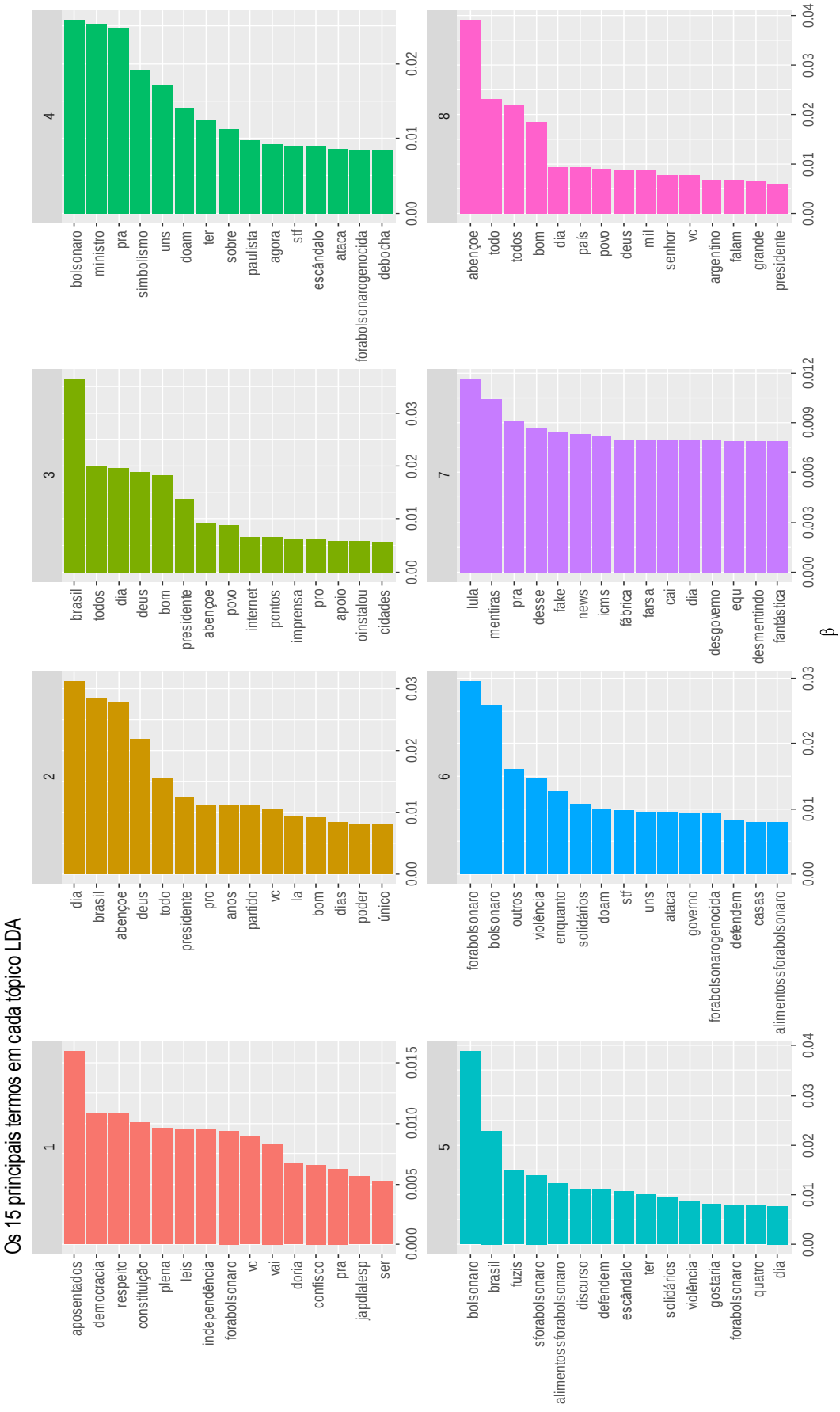
	topic	term	beta
1	1	aposentados	0.015971284
2	1	democracia	0.010853995
3	1	respeito	0.010832431
4	1	constituição	0.010079237
5	1	plena	0.009525521
6	1	leis	0.009501988
7	1	independência	0.009484125
8	1	forabolsonaro	0.009393202
9	1	vc	0.008924458
10	1	vai	0.008243608
11	1	doria	0.006684944
12	1	confisco	0.006538683
13	1	pra	0.006212600
14	1	japdlalesp	0.005602375
15	1	ser	0.005234287

Showing 1 to 16 of 120 entries

Fonte: Elaborado pelo autor.

Após obter o valor de β, conseguimos visualizar as principais palavras que compõem em um assunto, vamos definir esse conjunto de palavras como sendo um tópico, os 8 principais tópicos e suas respectivas palavras podem ser visualizados na Figura 17, juntamente as estimativas de β.

Figura 17: Gráfico de barras dos 15 termos com maior β para cada tópico.



Fonte: Elaborado pelo autor.

Vemos na Figura 17 as quinze principais palavras que compõem cada tópico, sendo elas as palavras com as maiores estimativas de beta, levando isso em consideração podemos notar que as estimativas de β são valores pequenos, isso significa que muitos termos pertencem a um tópico, ou seja, cada assunto discutido nos perfis é descritos com uma gama consideravelmente grande de palavras dado que as estimativas da distribuição de palavras por tópico é razoavelmente pequena.

Agora que conhecemos as estimativas de β, podemos calcular as estimativas de α referentes a distribuição de tópicos por documento, ou seja, qual tópico está fortemente associado a um documento. Temos então as seguintes estimativas para α, conforme o quadro 4.

Quadro 4: Estimativas de α.

		Candidato			
		A	**B**	**C**	**D**
Tópicos	**1**	0,9990388797	0,0000002323	0,0000002276	0,0000002607
	2	0,0000813180	0,0771593762	0,0000103191	0,0087011126
	3	0,0000002625	0,7617046065	0,0000002276	0,0000002607
	4	0,0000002625	0,0000002323	0,1281349544	0,0000002607
	5	0,0000002625	0,0000002323	0,4185007763	0,0000002607
	6	0,0000002625	0,0000002323	0,4533530395	0,0000002607
	7	0,0000002625	0,0000002323	0,0000002276	0,9912828077
	8	0,0008784897	0,1611348556	0,0000002276	0,0000147764

Fonte: Elaborado pelo autor.

Com objetivo de dizer qual candidato tem a maior probabilidade de estar associado a um assunto, podemos notar que o Candidato A apresenta aproximadamente 99% de chances de ser citado no assunto (tópico) um, enquanto o candidato B tem aproximadamente 76% de chance de estar ligado ao assunto três, já o candidato C possui 45% de chances de ser mencionado no assunto seis e 41% de chances no assunto cinco e por fim o Candidato D tem 99% de chances de estar ligado ao assunto sete.

7 Considerações Finais

Levando em conta as análises apresentadas temos a percepção de que os perfis dos candidatos são compostos em sua maioria por assuntos relacionados ao governo federal, entretanto através das análises conseguimos distinguir os candidatos e seus termos mais frequentes e relevantes.

As nuvens de palavras nos permitiram identificar quais são os termos mais repetidos no perfil de cada candidato e a proporção que eles são mencionados quando comparados a outros termos. Na análise de sentimentos conseguimos notar que os candidatos A, B e D possuem mais palavras positivas associadas a eles do que negativas, enquanto o candidato C se mostra mais associado a palavras identificadas como negativas. Entretanto, o método utilizado não reconhecia uma quantidade considerável de termos para que os resultados pudessem ser considerados significativos, mas sua aplicação em dados reais serviu como aprendizado e pode futuramente servir base para estudos direcionados a análise de sentimentos.

Os dendrogramas de uma forma simples e visual ajudaram a identificar a quantidade de assuntos presentes em todas as contas e o cálculo TF-IDF nos mostrou as palavras que são relevantes para cada candidato levando em conta os outros candidatos. Considerando que todos os perfis têm como principal assunto os acontecimentos relacionados ao governo o TF-IDF nos mostra termos que realmente diferenciam cada candidato e reforçam a análise de distribuição de tópico por documento do algoritmo LDA.

Por fim por meio da aplicação do algoritmo LDA, verificamos quais assuntos tem maior probabilidade de serem citados na conta de cada candidato, e quais palavras são associadas com maior probabilidade a cada tópico e através das estimativas podemos concluir que, devido ao caráter político dos perfis, os assuntos estão significativamente sobrepostos e misturados e é grande a quantidade de palavras que compõem um assunto considerando que muitas pessoas em todos os perfis discutem o mesmo assunto. Uma análise mais detalhada dos fatos históricos é capaz de nos trazer maiores informações sobre o contexto politico que envolve os textos, contudo esse estudo demanda a necessidade de um especialista na área, para conclusões acuradas do tema.

Considerando todos os pontos apresentados, podemos dizer que as técnicas aplicadas são capazes de evidenciar informações úteis em bancos de dados não estruturados como os documentos de texto.

8 Referências

AIZAWA, A. **A information-theoretic perspective of tf-idf measures**. Pergamon, Tokyo, 101-8430, ed. 39, p. 45-65, 2001-2002.

BERRY, M. W; KOGAN, J. **Text Mining**: Application and Theory. Tennessee: Wiley, 2010. Disponível em: https://books.google.com.br/books?hl=pt-BR&lr=&id=u-SrKyUrafsC&oi=fnd&pg=PR5&dq=info:HzxGhQBTK0wJ:scholar.google.com/&ots=WNnDqLxiH_&sig=NlkGhZEPqNRywnPYzQibBF-biyw#v=onepage&q&f=false
Acesso em: 30 agost. 2021.

Bishop, C. M; **Pattern Recognition and Machine Learning.** Springer, 2006.

BLEI, D. M. **Introduction to probabilistic topic models**. Communications of the ACM, 2011.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. **Journal of Machine Learning Research**, v. 3, n. Jan, p. 993–1022, 2003.

BUSSAB, W. O.; MIAZAKI, E. S.; ANDRADE, D. A. **Introdução à análise de agrupamentos**. In: SIMPÓSIO NACIONAL DE PROBABILIDADE E ESTATÍSTICA, 9, 1990, São Paulo. Minicurso [...]. São Paulo: Associação Brasileira de Estatística, 1990.

DE PAUL JOURNAL OF RESEARCH (DJSR). New York: Mani, K. P., Out. 2018. ISSN 2394-4412. Disponível em: https://scholar.google.com.br/scholar?q=Text+Mining:+Techniques,+Applications+and+Issues.&hl=pt-BR&as_sdt=0&as_vis=1&oi=scholart#d=gs_qabs&u=%23p%3Do9WwDzOZk78J
Acesso em: 18 agos. 2021.

GEMAN, S.; GEMAN, D. **Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images**. IEEE Transactions on Pattern Analysis and Machine Intelligence, Taylor & Francis, v. 6, n. 6, p. 721–741, nov. 1984.

GRIFFITHS, T. L.; STEYVERS, M. **Finding scientific topics**. PNAS, v. 101, n. suppl. 1, p. 5228–5235, 2004.

HEARST, M. **Untangling text data mining**. University of California, Berkeley, 1999, ed. 37.

LIU, B; CHENG, J. **Opinion observer: analyzing and comparing opinions on the Web**. In: International conference on word wide, 2012. p. 342-351.

MIGUEL, A. L. **Aplicaciones del text mining em la actualidad centrado em el área de marketing**. 2019. Trabalho de conclusão de curso – Faculdade de ciências econômicas y empresariales, Universidade Pontificia.

MORAIS, E.A.M; AMBRÓSIO, A.P.L. **Mineração de Texto. 2007**. Relatório técnico – Instituto de informática, Universidade Federal de Goiás.

Nasukawa, T. and Yi, J. **Sentiment Analysis**: Capturing Favorability Using Natural Language Processing Definition of Sentiment Expressions. 2003. 2nd International Conference on Knowledge Capture, p. 70–77.

PALAZZO, M. D. O; LOH, S; AMARAL, L. A; WIVES, L. K. **Descoberta de conhecimento em textos através da análise de seqüências temporais.** In: Workshop em algoritmos e aplicações de minaração de dados. Florianopolis: Sociedade Brasileira de Computação, volume II, p. 49–56, 2006.

RIVADENEIRA, A.W; GRUEN, D. M; MULLER, M. J; MILLEN, D. R. **Getting our head in the clouds: toward evaluation studies of tagclouds**. In: Conference on Human Factors, 2007. Anais [Computing Systems]. p. 995-998.

ROSA, J. **O significado da palavra para o processamento de linguagem natural**. In: Estudos Linguísticos XXLV, 1998, Campinas. Anais [Seminários do GEL]. Rio preto: UNESP-IBILCE, 1998. p. 7.

RUSSELL, S. J.; NORVIG, P. **Artificial Intelligence: A Modern Approach**. Pearson Education, 2003. ISBN 0137903952.

SILGE, J; ROBINSON; D. **Text Mining with R**. Califórnia: O'Reilly, 2017.

SILVESTRE, M.R. **Análise de agrupamentos aplicada a redes neurais mlp**. 2006. 108f. Relatório de Pesquisa - Faculdade de Ciências e Tecnologia, Universidade Estadual Paulista "Júlio de Mesquita Filho", Presidente Prudente, 2006.

SIVIC, J.; RUSSELL, B. C.; EFROS, A. A.; ZISSERMAN, A.; FREEMAN, W. T. **Discovering objects and their location in images.** In: IEEE International Conference on Computer Vision. [S.l.: s.n.], 2005.

SULLIVAN, D. **The need for text mining in business intelligence.** In: RAISINGHANI, M. Business intelligence in the digital economy. Dallas: ed. IGP, 2004. cap. 6, p. 98-110.
Disponível em:
https://scholar.google.com.br/scholar?q=sullivan+2000+text+mining&hl=ptBR&as_sdt=0&as_vis=1&oi=scholart#d=gs_qabs&u=%23p%3DySX89bsFvLMJ
Acesso em: 13 agos. 2021.

TALIB, R; HANIF, K; AYESHA, S; FATIMA, F. **Text mining: Techniques, applications and issues**. Internacional journal of advanced computer Science and applications, Paquistão, v.7, 11 nov. 2016, Computer Science, p. 415-418.

THURAISINGHAM, B. **Data mining: technologies, techniques, tools, and trends**. Florida: ed. CRC Press, 1999.