Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Análise de Sentimento usando Abordagem Léxica de Discursos do Senado Federal

Luis Filipe Campos Cardoso

Monografia apresentada como requisito parcial
para conclusão do Curso de Engenharia da Computação

Orientador
Prof. Dr. Thiago de Paulo Faleiros

Brasília
2022

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Análise de Sentimento usando Abordagem Léxica de Discursos do Senado Federal

Luis Filipe Campos Cardoso

Monografia apresentada como requisito parcial
para conclusão do Curso de Engenharia da Computação

Prof. Dr. Thiago de Paulo Faleiros (Orientador)
CIC/UnB

Prof. Dr. Luís Paulo Faina Garcia
Universidade de Brasília

Prof. Dr. Vinícius Borges
Universidade de Brasília

Prof. Dr. João Gondim
Coordenador do Curso de Engenharia da Computação

Brasília, 10 de maio de 2022

Dedicatória

Dedico esse trabalho a minha família pelo apoio e carinho. Ao meu pai (*In Memorian*), por ter sido minha maior referência de vida. Aos meus amigos, pelos momentos inesquecíveis que passamos juntos. E ao Alan Turing, por ser uma referência e uma inspiração para mim na área da Computação.

Agradecimentos

Agradeço ao meu orientador, prof. Dr. Thiago, pelo apoio dado durante a realização deste trabalho.

Resumo

A Análise de Sentimento é uma área da Mineração de Textos e do Processamento de Linguagem Natural que busca realizar a classificação de textos em linguagem natural, a fim de saber, de forma automática, o sentimento do texto. Por exemplo, saber se seu conteúdo expressa o sentimento positivo ou negativo.

Com isso, este trabalho apresenta uma análise de sentimento de discursos do Senado Federal usando a abordagem léxica. Essa abordagem usa recursos léxicos (dicionários) para classificar os textos. Dessa forma, foram escolhidos três dicionários para fazer essa classificação. Além disso, foi construído um novo dicionário em português do Brasil baseado em um dicionário já existente em língua inglesa para fazer uma comparação com os demais dicionários.

Este trabalho tem como objetivo aplicar a Análise de Sentimento usando abordagem léxica em discursos do Senado Federal e iniciar uma investigação nesta área aplicada ao objeto estudado, no caso os discursos. Assim, busca-se entender se essa técnica computacional é eficaz nos discursos analisados e se há abertura para aprimoramentos no futuro. Isso porque, não há rótulos que apresentam o sentimento desses discursos. Esses rótulos podem servir de objeto de avaliação por especialistas na área de política em suas pesquisas, pois os discursos podem ser utilizados como objeto de estudo para encontrar conflitos entre os parlamentares e outros padrões. Cabe destacar que o volume de discursos é alto e cresce a cada dia. Dessa forma, esse trabalho analisa discursos de senadores e senadoras da República de forma automatizada usando análise de sentimento usando dicionário para apoiar pesquisas nessa área.

Palavras-chave: Análise de Sentimento, discursos políticos, Senado Federal, Processamento de Linguagem Natural

Abstract

Sentiment Analysis refers to the Text Mining and Natural Language Processing area. Its goals seek to classify texts in natural language, in order to know their feeling automatically. I.e., if their content is positive or negative.

This work presents a sentiment analysis applied in Brazilian Federal Senate speeches adopting the lexical approach. This approach uses lexical resources (dictionaries) to classify sentiment's texts. Thus, three dictionaries were chosen to make this classification. In addition, a new dictionary in Brazilian Portuguese was built based on an existing dictionary in English Language to compare with other three dictionaries.

This work aims to apply Sentiment Analysis using a lexical approach in Brazilian Federal Senate speeches. This initiative explore this area applied to the subject of this study (senator's speeches). Therefore, we seek to understand whether this computational technique was effective in this scenario and whether there is a space for improvements in the future. Even because, there are no labels for these speeches at present. These labels can used by other researchers and political specialists in their analysis, as the speeches can be used as an object of study to find conflicts between parliamentarians and other patterns. It should be noted that the volume of speeches is high and grows every day. Thus, this work analyzes speeches by senators of the Republic in an automated way using sentiment analysis to support research in this area.

Keywords: Sentiment Analysis, political speech, Federal Senate, Natural Language Processing

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

INTRODUÇÃO

Na atividade parlamentar, os senadores e deputados fazem discursos na tribuna sobre diversos temas que vão desde de propostas e reivindicações até homenagens, votos em algum projeto de lei e opiniões sobre determinado assunto que impacta a sociedade. Com isso, o Senado Federal disponibiliza em seu repositório institucional a transcrição dos discursos realizados pelos parlamentares para que a sociedade tenha acesso a esse material (vide [2]). Mesmo porque, os órgãos públicos costumaram a disponibilizar seus dados ao cidadão, principalmente após a promulgação da Lei de Acesso a Informação (LAI - Lei 12.527 de 18 de novembro de 2011 [1]).

Dessa forma, os pesquisadores e a própria sociedade podem saber mais sobre o trabalho do Congresso Nacional, acompanhar as discussões que estão ocorrendo ou que já ocorreram e fazer diversas análises sobre os temas abordados. No entanto, tem um problema: há um volume significativo de discursos, o que pode dificultar na realização das análises, já que seria dispendioso ler e analisar todos os discursos manualmente. Sendo assim, uma alternativa é a utilização de recursos tecnológicos e computacionais, a fim de auxiliar a pessoa a qual está realizando a análise, como visto no artigo [14].

Uma das análises que pode ser realizada usando os discursos é verificar a presença de conflitos entre os grupos políticos. Geralmente, a respeito de algum tema divergente, a base governista costuma ter posições contrárias da oposição, como visto em [14]. Ou seja, quando a base governista traz um tema para a discussão, é muito provável que a oposição faça alguma crítica e mostre pontos contrários sobre o tema que foi levantado. Proskch e outros autores em seu artigo [14] mencionam que essa dinâmica pode ser vista em votação sobre o orçamento, por exemplo, onde a base governista tenta aprovar sua proposta e a oposição faz críticas: ou por não concordar com o orçamento proposto, ou para incluir na proposta suas demandas.

Neste contexto, conforme [10], um dos recursos computacionais disponíveis dentro da área de Processamento de Linguagem Natural é a Análise de Sentimentos. Ela tem sido

muito utilizada pela academia e pela indústria para realizar pesquisas, onde precisam ser analisadas opiniões de grande massas de dados, como análises de produtos, comentários de filmes, comentários em redes sociais, rastreamento de entidades e análises de sumários.

Ainda de acordo com Brito em [10], cabe mencionar que as empresas privadas, as instituições públicas e as instituições de ensino têm que lidar diariamente com um grande número de informações em suas atividades e na própria Internet, tais como notícias, normas legais, relatórios, dentre outros documentos. Esse panorama é um indicativo para a construção de ferramentas que auxiliem o processamento e a análise de sentimentos na língua portuguesa para facilitar a compreensão dessas informações, visto que 80% das informações armazenadas nas empresas são produzidas de forma não estruturada e 80% do conteúdo contido na Internet está em formato textual.

Com o Processamento de Linguagem Natural é possível recuperar informações, extrair dados, resumir textos, descobrir padrões, associações e regras e realizar análises qualitativas e quantitativas em documentos. Com isso, há pesquisas na área de processamento de linguagem natural em universidades de todo o mundo nas mais variadas línguas. No entanto, há muito que ser feito nessa área, as pesquisas em mineração de textos remetem a uma forte regionalização, apresentando maiores resultados para algumas línguas do que para outras dada as características de cada uma [10].

Existem pesquisas dessa área para a língua portuguesa, mas os resultados ainda são tímidos. Boa parte dos resultados concentram-se na língua inglesa, e estes não são, na maioria das vezes, transportáveis para outras línguas [10]. Conforme enunciado por Sousa em [15], tem um esforço da comunidade acadêmica para a criação de recursos léxicos, como dicionários, para a língua portuguesa. São exemplos: OpLexicon, SentiLex, Brazilian Portuguese Linguistic Inquiry and Word Count (LIWC) e Onto.PT.

Segundo Benevenuto em [9],o objetivo principal da Análise de Sentimentos é definir métodos capazes de extrair informações subjetivas de textos em linguagem natural de forma automática, como opiniões e sentimentos, a fim de gerar conhecimentos estruturados que possam ser úteis para a tomada de decisão. Devido a importância dessa área, esse tema abrange não só a computação, como também a psicologia e as ciências sociais. Essa técnica pode ser usada, por exemplo, em trabalhos para o ramo da Ciência Política, como é o caso do trabalho visto no artigo [14].

Balage em seu artigo [8] explica que a Análise de Sentimento, ou a Mineração de Opinião, ganhou muita força com o crescimento das redes sociais na Internet e uma aplicação comum é para a classificação de textos (classificar a opinião como positiva, negativa ou neutra). Há duas abordagens de Análise de Sentimento: usando aprendizado de máquina e baseada em léxico. A abordagem de aprendizado de máquina utiliza um conjunto de recursos (por exemplo, vocabulário) que é treinado a partir de uma corpora

ou de exemplos rotulados. A abordagem baseada em léxico usa dicionários ou grupos de palavras pré-selecionadas para fornecer a polaridade, ou orientação semântica, para cada palavra ou frase no documento analisado. A abordagem baseada em léxico não requer, a princípio, de exemplos rotulados e é conhecida por sua independência de domínio, enquanto a abordagem de aprendizado de máquina tende a se adaptar ao domínio em que o classificador foi treinado [8].

Este recurso poderia verificar a opinião de cada parlamentar ou grupo político de forma automática em algum recorte temporal ou tema de discussão. Dessa forma, poderia ser possível verificar se o parlamentar ou grupo de parlamentares, ao realizar seu discurso, tem uma opinião positiva ou negativa. Para isso, é importante verificar se há algum método da Análise de Sentimento que consiga trazer essa informação sobre os discursos dos senadores de forma confiável.

Este trabalho de conclusão de curso tem como objeto analisar o sentimento de discursos de senadores do congresso nacional brasileiro, usando uma abordagem léxica e quatro diferentes dicionários (sendo um dos dicionários criado neste projeto). A finalidade desta análise é classificar os discursos em positivo e negativo para que essa informação seja utilizada para facilitar pesquisas, onde o objeto de estudo sejam os discursos de parlamentares.

Com base em trabalhos similares sobre Análise de Sentimento, foi estabelecida uma hipótese para a pesquisa, já que é uma questão pertinente ao contexto que está sendo estudado. Dessa forma, a hipótese é: *é possível classificar os discursos em positivos e negativos, mesmo eles possuindo tamanhos variados e, em muitos casos, sendo maiores que tweets.*

1.1 Objetivo

O objetivo geral deste trabalho de conclusão de curso é construir um dicionário léxico para a classificação de sentimento (positivo ou negativo) de discursos de senadores(as) da República usando a abordagem léxica e comparar com dicionários já existentes.

Para alcançar esse objetivo geral, foram determinados os seguintes objetivos específicos:

- estudar técnicas de análise de sentimento disponíveis na literatura;
- criar um dicionário para analisar o sentimento dos discursos;
- encontrar dicionários prontos que analisem os mesmos discursos;
- avaliar os resultados das análises de sentimentos dos discursos para cada dicionário usado neste trabalho e avaliar a hipótese desse trabalho.

1.2 Organização da monografia

O primeiro capítulo apresenta uma introdução ao tema deste trabalho e uma introdução ao contexto do objeto, além de fazer um descrição do problema apresentado neste trabalho. Ato contínuo, o segundo capítulo mostra a fundamentação teórica deste projeto, onde foram descritos conceitos relacionados aos temas deste estudo e foram analisados trabalhos similares, além de apresentar a área estudada e as ferramentas utilizadas. O terceiro capítulo expõe o desenvolvimento do trabalho, detalhando as técnicas utilizadas e a metodologia aplicada a ele. Já no quarto capítulo, são divulgados os resultados do trabalho e uma análise crítica sobre ele. Por fim, o capítulo cinco apresenta a conclusão deste estudo, enuncia o que foi aprendido durante o trabalho e expões possíveis futuras investigação a respeito do tema.

Capítulo 2

FUNDAMENTAÇÃO E REVISÃO E BIBLIOGRÁFICA

2.1 Mineração de texto

Pezzini explica em [13] que a Mineração de Textos foi criada para resolver problemas como: entender a linguagem natural dos documentos de texto e conseguir lidar com a sua imprecisão e incerteza. Essa área é composta de diversas áreas da computação, tais como: Mineração de Dados, Recuperação de Informação, Aprendizado de Máquina e Linguagem Computacional, e assim transformar o texto em algo que um computador possa compreender.

Um dos objetivos da Mineração de Textos é extrair conhecimento de grande volume de dados e criar conexões e padrões entre eles com base em certas métricas como frequência e temática dos termos encontrados. O processo de mineração não foi criado para ser um mecanismo de busca, já que ela procura informações que não são conhecidas de seus usuários. Mesmo porque, em mecanismos de busca, os(as) usuário(as) já sabem o que querem procurar. Além disso, a mineração não tenta simular o comportamento humano, por isso não é uma de suas aplicações desenvolver robôs de conversação (também conhecidos como *chatbots*) [13].

Evangelista em [11] define que, de forma geral, a Mineração de Textos envolve três etapas: pré-processamento, análise dos dados (ou extração) e avaliação de descobertas. A primeira etapa (pré-processamento) está relacionada a limpeza dos dados. Geralmente, é realizada a remoção das *stopwords* (artigo, rejeição, preposição ou palavras com pouco significado para o texto) e o *stemming* (remoção de variações de plural, gerúndio e sufixos temporais).

A etapa seguinte (análise dos dados ou extração) busca extrair conhecimento através das informações contidas nos textos analisados utilizando alguma técnica computacional.

A Mineração de Textos trabalha com a aplicação de algoritmos que processam grandes volumes de textos em linguagem natural, as quais, muitas vezes, não estão estruturados, com a finalidade de extrair informações relevantes, úteis e inovadoras não explícitas através da identificação de regularidades, tendências e de padrões não triviais. A última etapa (avaliação de descobertas) tem como objetivo avaliar os resultados e apresentar os achados.

2.2 Processamento de linguagem natural

Sousa explica em [15] que o Processamento de Linguagem Natural, ou PLN, é uma área de pesquisa da Inteligência Artificial e da Mineração de Textos que utiliza o computador para manipular, processar e compreender linguagem natural, ou seja, aquela escrita ou falada. O objetivo dessa área é reunir o conhecimento sobre como os humanos entendem a linguagem natural e, assim, criar técnicas e ferramentas computacionais para manipular e compreender a linguagem.

Como visto na introdução, a quantidade de textos existentes na Internet alcançou proporções não gerenciáveis. Por conta disso, tornou-se quase impossível fazer buscas e análises nesse grande volume de dados usando apenas comandos SQL (*Structured Query Language* - Linguagem de Consulta Estruturada), por exemplo. O ideal seria que as pessoas pudessem fazer perguntas em suas próprias linguagens, sem a necessidade de usar alguma linguagem de programação ou linguagem de máquina [15].

No entanto, o principal desafio para o avanço do Processamento de Linguagem Natural é a ambiguidade presente na linguagem humana. A ambiguidade pode tornar o processamento de linguagem natural complexo. Posto que ela pode tornar a maioria das técnicas desenvolvidas para análise de linguagens naturais ineficazes, quando aplicadas em textos com ambiguidade. A monografia [15] cita algumas das principais causas de ambiguidade na linguagem natural:

- **Ambiguidade lexical:** ocorre quando uma mesma palavra tem mais de um mesmo significado. Por exemplo, a palavra *cobre* pode ser referir tanto ao metal cobre, quanto ao verbo cobrir;
- **Ambiguidade sintática:** ocorre quando podem ser feitas diversas interpretações de uma mesma oração. Por exemplo, *Maria leu a notícia sobre a greve no campus.* Nesta oração, pode ser interpretado que Maria leu sobre a greve quando estava no campus ou leu que a greve ocorre no campus;
- **Ambiguidade pragmática:** ocorre quando não se sabe o significado da expressão na sentença ou sua intenção no momento da sua enunciação. Por exemplo, na

oração, *Paulo vai à escola* não tem como saber se Paulo é estudante ou se está indo à escola neste momento;

- **Ambiguidade semântica:** ocorre quando uma palavra possui mais de um conceito aplicado relacionado. Por exemplo, na oração, *um rio corre através de cada país europeu*, pode-se entender que um único rio corre através de todos os países europeus ou vários rios correm através de diversos países europeus;
- **Ambiguidade predicativa:** ocorre quando há algum problema relacionado a interpretação textual. Por exemplo, na frase, *a crítica deste autor* não tem como saber se o autor é objeto da crítica ou quem faz a crítica.

2.3 Análise de sentimento

Sousa informa em [15] que os textos podem ser classificados em dois grandes grupos: fatos e opiniões. Os fatos representam expressões objetivas sobre determinado assunto, pessoa, serviço, objeto, dentre outros. Por outro lado, as opiniões estão relacionadas a sentimentos, avaliações, atitudes e emoções, as quais tem caráter subjetivo. Nesse contexto, a Análise de Sentimento pode ser compreendida como uma área que estuda e analisa essas opiniões. Suas técnicas são utilizadas para extrair essas opiniões e determinar de forma automática os sentimentos presentes em linguagem natural.

Neste cenário, as opiniões podem estar escondidas em páginas de sites, blogs e postagens em redes sociais, por exemplo. Por conta dessa variedade de locais e volume de dados, essa tarefa de procurar informações relevantes e extrair opiniões de textos pode ser muito complicada se feita manualmente. Neste ponto, a Análise de Sentimentos feita de forma automatizada se faz necessária. Ela pode ser aplicada em análises de bolsa de valores, notícias jornalísticas, debates políticos, por exemplo [15].

Um dos objetivos da Análise de Sentimentos é criar técnicas capazes de extrair informações subjetivas, de forma automática, de textos em linguagem natural, com a finalidade de gerar conhecimento estruturado que possa subsidiar alguma análise ou um sistema de tomada de decisão. A Análise de Sentimentos é uma das áreas de pesquisa do Processamento de Linguagem Natural. Suas aplicações são variadas e podem apoiar tanto empresas para saber o sentimento dos usuários de seus produtos (muito usado no Marketing e em Relações Públicas), quanto pesquisadores das mais variadas áreas do conhecimento, como Psicologia, Ciências Sociais e Linguística [15].

A Análise de Sentimentos em documentos não é uma atividade trivial, pois uma busca simples por palavras-chave - como “bom” e “ruim” - não é suficiente para determinar o sentimento expresso no texto. Por exemplo, na oração *Ela não é um pessoa boa*, a presença

da palavra *boa* não torna a sentença positiva. Dessa forma, há a necessidade de se utilizar de técnicas mais avançadas para analisar o sentimento dos textos [15].

De modo geral, essa área pode trazer informações relacionadas à orientação semântica ou polaridade das palavras ou expressões contidas no texto. Neste cenário, as expressões podem ser categorizadas pela sua polaridade, onde são comumente representadas como expressões negativas, positivas e neutras. Essa última categoria pode ser ou não adotada e ela representa os casos em que o sentimento associado a uma certa expressão não pode ser declarado de forma objetiva como positiva ou negativa, sendo assim necessário um contexto (por exemplo, os juros tiveram uma queda importante) [15].

Cabe ressaltar que a Análise de Sentimento trata de problemas de classificação. Com isso, para fazer essa classificação, ela busca diferenciar textos de acordo com a sua polaridade, mesmo que uma frase não denote explicitamente um sentimento. Quando se descreve um fato, uma oração pode ser classificada como positiva ou negativa para uma determinada área, por exemplo a saúde e políticas públicas [10].

Junqueira aborda em [12] que para analisar o sentimento dos textos, podem ser utilizadas duas abordagens: léxica e aprendizado de máquina. A abordagem através do aprendizado de máquina cria um modelo de classificação que usa associação de mensagens a uma polaridade ou um sentimento. Para criar um modelo de classificação, alguns métodos utilizam dados para treinamento e esses dados de treinamento são constituídos por mensagens de opinião, onde cada mensagem está associada a uma sentimento ou polaridade. Após o treinamento, é possível fazer a associação dos textos ainda não classificados, tendo por base os textos que possuem classes associadas.

Geralmente, utilizam-se métodos supervisionados, os quais são dependentes de um treinamento prévio. Com isso, é construído um modelo conciso para realizar predições. Podem ser usadas várias técnicas como Naïve Bayes, Redes Bayesianas, Redes Neurais Artificias, SVM, árvores de decisão e modelos baseados em regras[12].

2.4 Análise de sentimento usando abordagem léxica

Diferentemente da abordagem usando aprendizado de máquina, a Análise de Sentimento usando Abordagem Léxica usa dicionários ou coleções de palavras de opinião com anotações que indicam o sentimento em positivo ou negativo. Essa técnica pode ser construída de forma manual ou automática através de uma lista de palavras. Em alguns casos, são utilizados adjetivos para orientar a semântica do texto e eles são listados e compilados em um dicionário. Assim, os adjetivos e termos dos textos são extraídos e usados para pontuá-los, a fim de orientar a polaridade do texto[12]. Para fazer a classificação se-

mântica, é levada em consideração a relação de uma determinada palavra com os demais termos do texto [15].

A Análise de Sentimento usando Abordagem Léxica se divide em duas abordagens: dicionário e corpus. Ao usar dicionários, são utilizados um conjunto de palavras com anotações de polaridade para classificar a sentença de acordo com a polaridade de cada palavra do dicionário [12]. São exemplos de dicionários no idioma português: SentiLex[5], OpLexicon[4] e Onto.PT[3]. Cabe ressaltar que a utilização de um bom recurso léxico (dicionário) é essencial para que a técnica tenha sucesso [15].

Na abordagem léxica, um dicionário de palavras (ou recurso léxico), ao invés de possuir conteúdo ou significado para cada uma das palavras, tem em seu lugar um significado quantitativo (por exemplo, um número entre -1 a 1, onde -1 é atribuída a uma palavra negativa e 1 para a uma positiva) ou mesmo um valor qualitativo (por exemplo, positivo/negativo e feliz/triste). Nesta técnica, é assumido que as palavras individualmente possuem uma polaridade prévia, ou seja, uma orientação semântica (ou polaridade) independente de qualquer tipo de contexto e, assim, expressar essa polaridade com um valor numérico ou classe [9].

A outra abordagem é a baseada em Corpus. Uma característica dessa abordagem é a sua dependência com padrões sintáticos ou padrões que ocorrem no corpo do texto. Pode-se definir Corpus como o corpo do texto de linguagem natural usado para acumular algum tipo de estatística sobre a linguagem natural. As informações extraídas desse método podem ser anotações de cada palavra que indicam suas função dentro do texto, por exemplo, se é um adjetivo, substantivo, advérbio, ou outro. Essa mesma abordagem pode ser dividida em baseada em estatísticas e baseada na orientação semântica [12]. Essa técnica costuma usar grandes corpora linguísticos com o objetivo de definir para cada palavra um valor estático, o qual determinará sua polaridade. Sendo assim, pode-se dizer que técnicas baseadas em corpus podem não ser tão eficientes quanto às técnicas baseadas em dicionário [15], já que pode utilizar mais recursos computacionais e a abordagem léxica tem uma capacidade maior de predição se comparada com a outra abordagem [9]. Por isso, foi usada a abordagem léxica usando dicionário neste trabalho.

Um dos conceitos abordados acima foi o de polaridade. A fim de firmar seu entendimento, a polaridade pode ser entendida como o grau de positividade ou negatividade de um texto. Geralmente, ela é a saída ao executar um dos métodos de Análise de Sentimentos abordados acima. Certos métodos tratam a polaridade como um resultado discreto binário (positivo ou negativo) ou ternário (positivo, negativo ou neutro). Neste sentido, a oração *O dia está bonito hoje* é positiva e a oração *O trabalho de hoje foi péssimo* é negativa, já a oração "Amanhã é 21 de Abril" não possui polaridade e costuma ser classificada como neutra [9].

2.5 Trabalhos relacionados

Para a realização deste trabalho de graduação, foram lidos artigos e outros tipos de textos acadêmicos que se relacionavam de alguma forma com este projeto. Com isso, foram escolhidos trabalhos na área de Análise de Sentimento sob os pontos de vista da área da Computação e da Ciência Política, já que a primeira é a área principal deste trabalho e a segunda é a área do objeto de estudo escolhido (discursos dos(as) Senadores(as) da República).

Sendo assim, seguem resumos de alguns dos trabalhos relacionados a esta monografia em tela:

- **[12]:** esse trabalho analisa o sentimento de tweets relacionados às Olimpíadas de 2016 usando a abordagem léxica (com os dicionários: SentiLex, OpLexicon e LIWC) e algoritmos de aprendizado de máquina (com os algoritmos: *Naïve Bayes*, SVM, Máxima Entropia, *Random Forest* e Árvore de Decisão). O algoritmo SVM teve o melhor desempenho. Os autores usou algumas métricas para comparar as abordagens e algoritmos: Precisão (mostra a quantidade de positivos e negativos acertados), Acurácia (mede o desempenho das abordagens considerando os registros corretamente classificados em relação a todos os registros), *Recall* (apresenta a relação entre os registros classificados pertencentes a determinada classe, positivo e negativo, em relação ao total de registros de cada classe), Abrangência (apresenta a relação entre as palavras contidas nas mensagens de opinião e as palavras do dicionário e, assim, é medir os termos comuns entre os dicionários e as mensagens classificadas) e Medida de Concordância (é o coeficiente Kappa usado para definir um limiar de concordância entre os anotadores em tarefas de classificação). O trabalho foi executado em sete etapas: coleta de dados, armazenamento dos dados, seleção, treinamento e classificação manual, pré-processamento, mineração e análise dos resultados;

- **[13]:** esse trabalho explica conceitos, o processo da mineração de textos e suas aplicações e técnicas;

- **[14]:** esse trabalho faz um experimento usando o dicionário *Lexicoder Sentiment Dictionary* - Dicionário de Sentimentos Lexicoder (LSD), onde os pesquisadores analisaram o sentimento de discursos de congressistas dos mais variados países. Eles traduzem o dicionário para outras línguas para poder analisar discursos em outras línguas, inclusive em português de Portugal. Eles buscam entender se é possível analisar o sentimento através de um dicionário de sentimento específico para política e que pudesse ser traduzido para outras línguas. Uma das análises que

eles fizeram com a análise de sentimento dos discursos é a identificação de possíveis conflitos entre grupos políticos;

- **[10]:** o objetivo desse trabalho foi implementar um modelo de sistema para classificar automaticamente sentimentos em bases textuais escritas em português do Brasil, utilizando os conceitos da aprendizagem de máquina. O autor usa técnicas para limpeza dos discursos, como a remoção de *stopwords* e como a normalização de palavras (com o *stemming*). Para analisar a performance dos classificadores, foi usada uma matriz de confusão, dentre outras ferramentas;
- **[9]:** esse minicurso oferece uma introdução à Análise de Sentimento. É apresentada uma visão geral sobre o tema, são explicados conceitos (como polaridade, força do sentimento, dentre outros) e suas aplicações mais populares, abordando técnicas supervisionadas e não supervisionadas. Além disso, é explicado o processo de construção de dicionários, os quais são utilizados na abordagem léxica; bem como exemplos de dicionários disponíveis;
- **[15]:** esse trabalho apresenta a implementação de classificadores semânticos que realizam uma análise de sentimentos em textos escritos em português e utiliza o recurso léxico SentiWordNet, traduzido de forma automática para o português, e compara o seu desempenho em relação ao SentiLex, recurso léxico já em português. O autor explica as abordagens da análise de sentimento e seus algoritmos, métodos e técnicas. Além disso, ele explica os algoritmos utilizados para realizar a análise de sentimento usando a abordagem léxica.

Capítulo 3

ANALISADOR LÉXICO DE SENTIMENTOS DE DISCURSOS POLÍTICOS

3.1 Linguagem R

Neste trabalho, foi usada a linguagem de programação R. Ela pode ser definida como uma linguagem multi-paradigma orientada a objetos ou funcional e é fracamente tipada. É usada principalmente para análise e visualização de dados e foi criada por Ross Ihaka e por Robert Gentleman (da Universidade de Auckland - Nova Zelândia).

O código fonte do R é disponibilizado sob a licença GPL e funciona em diversos Sistemas Operacionais (Windows, MacOS e Linux). Na instalação do R, os pacotes são instalados usando uma rede de distribuição do R (em inglês, CRAN - *Comprehensive R Archive Network* ou Rede Abrangente de Arquivos R). Nos últimos anos, essa linguagem se tornou bem popular, especialmente no desenvolvimento de análises estatísticas e de dados.

3.2 Quanteda

O Quanteda é um pacote da linguagem R usado neste projeto, criado por Kenneth Benoit, Kohei Watanabe e outros desenvolvedores suportado inicialmente pelo European Research Council (Conselho Europeu de Pesquisa). Esse pacote foi escolhido por estar no artigo [14], que inspirou parte do projeto. O pacote é usado para gerenciar e analisar textos como dado, através do uso do processamento de linguagem natural. A ferramenta exige o conhecimento da linguagem R, possui recursos de visualização e processamento de dados e foi projetada para ser eficiente em sua execução.

Para instalar o Quanteda, é necessário ter o R na máquina que vai processar os dados e suas funcionalidades podem ser usadas instalando o pacote *quanteda* disponível no CRAN. A documentação pode ser acessada através do link [7]. O Quanteda neste trabalho foi usado para tokenizar o texto e fazer as operações (principalmente, Document-Feature Matrix - Matriz do Documento-Termo - ou DFM) para classificar o sentimento de cada palavra do dicionário.

3.3 Etapas do trabalho

O trabalho foi realizado em seis etapas: escolha do objeto de estudo, escolha da área de pesquisa dentro da Inteligência Artificial, estudo de bibliografia correlata à área e ao objeto de estudo, criação de dicionário próprio, experimento com o dicionário próprio e outros disponíveis na academia e, por fim, análise dos resultados. Seguem o resumo de cada etapa:

1. **Escolha do Objeto de Estudo:** para realizar o trabalho de graduação, foi necessário escolher um objeto de estudo, ou seja, algum recurso a ser estudado durante a execução das atividades. Como os estudos para o desenvolvimento deste trabalho estavam associados a um grupo de pesquisa com discursos de senadores da República do Brasil, foi escolhido esse objeto, o qual ainda não foi totalmente explorado. Mesmo com todos os desafios que serão abordados ao decorrer desta monografia, foi constatado que seria um recurso que poderia gerar uma pesquisa interessante na área da Inteligência Artificial.

2. **Escolha da Área de Pesquisa:** ao analisar os discursos, percebeu-se que não havia muitos dados rotulados. A partir do artigo [14], percebeu-se que seria possível realizar um experimento na área de Análise de Sentimento usando os discursos de senadores do congresso brasileiro. O artigo em questão aborda a questão da Análise de Sentimento no contexto da Ciência Política, o qual é o mesmo contexto do objeto de estudo escolhido. Além disso, o artigo explora a questão do uso de um mesmo dicionário léxico para diversas línguas distintas. Diante deste cenário e pelo grupo de pesquisa ainda não ter realizado pesquisa de Análise de Sentimento usando os discursos, foi escolhida essa área, a fim de criar métodos para classificar os textos e, assim, criar mecanismos e rótulos para o discurso de forma automática, baseado em sua polaridade (positivo e negativo).

3. **Estudo da Bibliografia:** após a escolha da área de pesquisa - Análise de Sentimento - foram analisados diversos artigos similares, com a finalidade de entender o

que foi feito anteriormente em outras pesquisas. Esses trabalhos foram importantes para fomentarem ideias para implementar este projeto.

4. **Criação de Dicionário:** com base no artigo [14], foi construída uma versão do dicionário desse artigo citado, só que de língua portuguesa do Brasil. Os procedimentos e detalhes de construção desse novo dicionário serão explorados mais a frente.

5. **Experimento:** após a construção do dicionário, foi conduzido um experimento com 18.000 discursos de senadores da República, onde foi analisado seu sentimento usando o dicionário criado neste trabalho, além de outros três encontrados em trabalhos correlatos.

6. **Análise de Resultado:** por fim, após o experimento, foi feita uma análise dos resultados da Análise de Sentimento dos quatro dicionários experimentados neste projeto.

3.4 Análise do objeto de estudo: discursos de senadores(as)

Foram analisados, neste trabalho, 18 mil discursos do Senado Federal. Esses discursos, além de seu texto, contém meta-dados, os quais trazem informações complementares sobre eles. São exemplos de informações complementares dos discursos:

- Nome do(a) Senador(a);
- Partido Político do(a) Senador(a) na Época;
- Unidade da Federação do(a) Senador(a);
- Data do Pronunciamento;
- Resumo do Discurso;
- Indexação, dentre outros.

O discurso e seus conteúdos estão estruturados em formato JSON e armazenados em arquivos do tipo *pickle*. Essa estratégia foi usada para armazenar os objetos (do tipo *string*) que podem ser serializada e recuperada facilmente. Os textos dos discursos tem tamanhos variados, variando de 3 a 18537 palavras (aproximadamente após remoção das *stopwords* e alguns símbolos).

Tabela 3.1: Dados Estatísticos dos Discursos Analisados neste Trabalho.

Média Aritmética das Palavras dos Discursos	817,10
Número de Palavras Distintas	2.691,00
Primeiro Quartil - Quantidade de Palavras	319,80
Segundo Quartil - Quantidade de Palavras	650,50
Terceiro Quartil - Quantidade de Palavras	1.122,00
Menores Quantidades de Palavras dos Discursos	3, 4, 5, 6 e 7
Maiores Quantidades de Palavras dos Discursos	10579, 11798, 14400, 16134 e 18537

Seguem alguns dados estatísticos sobre os textos dos discursos na Tabela 3.1, após remoção das *stopwords* e alguns símbolos. Percebeu-se que os textos tem um perfil mais longo, se comparado a trabalhos que analisam sentimento de *tweets*, como é o caso do artigo [10], dado que a média é 817,10 palavras e o segundo quartil (mediana), 650,50 palavras. A partir desses dados são levantadas algumas questões e hipóteses, como: *é possível classificar os discursos em positivos e negativos, mesmo eles possuindo tamanhos variados e, em muitos casos, sendo maiores que tweets.*

Após analisar os discursos de forma ampla, viu-se a necessidade de analisar as palavras em si, com a finalidade de entender sua frequência e a frequência dessas palavras nos textos estudados. Conforme a Tabela 3.2, verificou que as palavras mais frequentes são: *país*, *hoje*, *fazer* e *nacional* e elas, em média, aparecem de 3 a 4 vezes por discurso. Com isso, a tabela citada apresentou as palavras mais frequentes nos discursos, a frequência delas em todos os discursos, a frequência de discursos que tem aquela palavra e a média de vezes que aparece aquela palavra por discurso.

Todos esses dados são estatísticos e podem ser usados por especialistas da área de política para criar suas análises e tirar suas conclusões. Além da tabela, há outros tipos de visualização, como é o caso da nuvem de palavras. A Figura 3.1 ilustra uma nuvem de palavra que apresenta as palavras mais vistas nos discursos de uma forma mais visual e mais amigável. Essa nuvem de palavras apresenta de forma clara, as oitenta palavras mais frequentes.

Todos esses recursos são relevantes para realizar análises específicas. Com isso, especialistas podem fazer recortes, a fim de trazer conclusões e insumos para suas pesquisas e relatórios. Apesar de entender que esse recurso pode ser usado de forma aplicada para especialistas, essa atividade trouxe um ponto muito importante, é possível analisar as palavras de um texto usando recursos tecnológicos. Entretanto, é essencial que se tenha um certo contexto e certos insumos que possam direcionar o trabalho para que os métodos computacionais aplicados sejam mais efetivos.

A partir dos pontos vistos acima, entende-se que seria possível explorar mais o uso das palavras nos textos utilizando recursos computacionais. Sendo assim, foi levantada

Tabela 3.2: Frequência das Palavras com Maior Frequência nos Discursos.

Rank	Palavra	Freq. da Palavra	Freq. dos Discursos	Média por Discurso
1	país	64506	13183	4.893120
2	hoje	53014	13299	3.986315
3	fazer	41235	11905	3.463671
4	nacional	37753	11169	3.380159
5	grande	35290	11040	3.196558
6	quero	33642	10288	3.270023
7	casa	31651	10248	3.088505
8	agora	29624	9842	3.009957
9	política	27879	8554	3.259177
10	pessoas	27849	8618	3.231492
11	bem	27293	10179	2.681305
12	projeto	26081	6981	3.735998
13	povo	25394	7564	3.357218
14	trabalho	25126	8185	3.069762
15	estados	23968	7816	3.066530
16	maior	23814	9705	2.453787
17	momento	23643	9743	2.426665
18	brasileiro	23606	8995	2.624347
19	tempo	23597	9687	2.435945
20	brasileira	23239	9148	2.540337

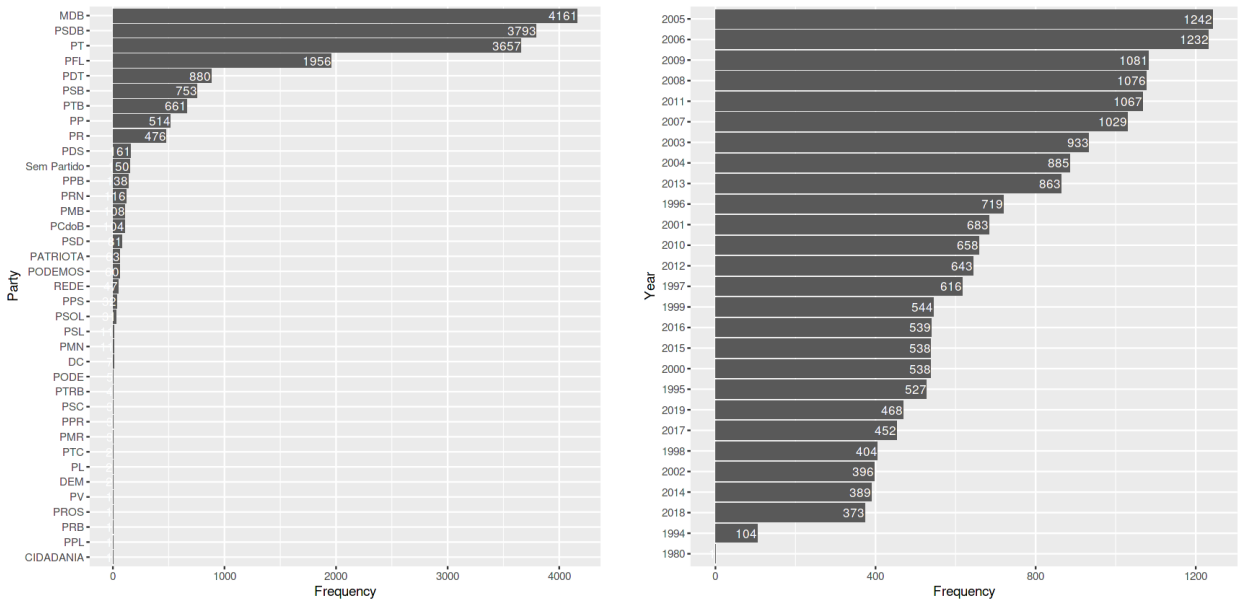
a hipótese se seria viável ou não analisar sentimento, a partir das palavras contidas no texto e de um recurso léxico (dicionário) previamente construído.

Com a finalidade de finalizar essa primeira análise dos discursos, colocou-se o foco nos metadados mencionados no início dessa seção. Dessa forma, percebeu-se que os discursos mais analisados foram dos partidos MDB (e PMDB), PSDB, PT, PFL e PDT, conforme a Tabela I.1 e a Figura (a) da Tabela 3.3. Além disso, foram analisados discursos de senadores e senadoras de 1980 a 2019, conforme a Tabela I.2 e a Figura (b) da Tabela 3.3, sendo de 2005, de 2006 e de 2009 a maioria dos discursos.

3.5 Criação do dicionário léxico

Para realizar a Análise de Sentimento dos discursos, neste projeto, foi construído um dicionário capaz de listar palavras que indicassem opiniões negativas e positivas. Indo a esse encontro, utilizou-se como base o dicionário *Lexicoder Sentiment Dictionary* (LSD) como referência, já que é um dicionário da língua inglesa especializado em política, o qual é a área dos discursos. Esse dicionário mencionado foi desenvolvido através do trabalho [14] e pode ser acessado através do Quanteda [7].

Figura 3.1: Nuvem de Palavras com Maior Frequência nos Discursos



(a) (b)

Tabela 3.3: Gráficos que mostram a frequência dos discursos analisados neste trabalho por partido e por ano

O problema é que os discursos analisados neste trabalho estão na língua portuguesa do Brasil e o dicionário está na língua inglesa. Uma ideia seria traduzir os discursos para o Inglês para poder utilizar o dicionário LSD. No entanto, seria um esforço computacional muito grande traduzir todos os discursos somente para se utilizar o LSD. Além disso, cabe ressaltar que cada língua possui expressões típicas, a qual só ela possui.

Essa temática é abordada no artigo [14] e ele sugere que, ao invés de traduzir os discursos, uma possível abordagem seria traduzir o dicionário e usar o dicionário traduzido para analisar o sentimento do discurso. Os resultados obtidos nesse referido artigo foram significativos e trouxeram um resultado positivo, mesmo que preliminarmente. Ou seja, houve bons resultados ao se traduzir o dicionário para diversas línguas e, em um dos experimentos realizados, a análise automatizada conseguiu acertar entre 75% e 80% dos casos na língua portuguesa (Portugal).

Diante do exposto, buscou-se fazer algo semelhante ao trabalho do artigo abordado, a fim de verificar se haveria algum resultado positivo ao fazer a análise do sentimento dos discursos dos(as) Senadores(as), utilizando esse dicionário traduzido. Com isso, foi realizada a tradução do LSD para o português do Brasil usando:

- a linguagem R;
- o pacote Quanteda [7];
- a solução de tradução da Microsoft (Microsoft Translator) através de sua API (*Application Programming Interface* ou Interface de Programação de Aplicativos, em português do Brasil), para fazer a tradução automática de cada palavra do dicionário LSD;
- a solução de tradução do Google (Google Tradutor).

Como o dicionário apresentava palavras em inglês *stemizadas*, foi necessário, antes de fazer a tradução automática, realizar a reversão do *stemming*, a fim de deixar as palavras com alguma flexão. Mesmo não sendo possível restabelecer as palavras para sua forma original especificamente, tratou-se de fazer a reversão para alguma forma flexionada para que ela pudesse ser traduzida automaticamente com mais facilidade. Esse ponto inclusive foi discutido no artigo [14], o qual usou abordagem similar.

Foi utilizada nesse processo a função *stemCompletion* do pacote *tm* da linguagem R para completar as palavras, pois o dicionário LSD armazena apenas seus radicais. Isso se deve ao fato de que, na hora de fazer a avaliação de sentimento da palavra, são avaliados seus radicais, porque uma palavra pode ter diversas flexões (gênero, número, grau...).

Para rodar a função *stemCompletion*, é essencial escolher um dicionário para completar as palavras. Com isso, deve-se mencionar o dicionário no parâmetro *dictionary*. Dessa

Tabela 3.4: Quantidade de termos em cada dicionário.

Polaridade/Dicionário	LSD em inglês	LSD Traduzido - português do Brasil
Positiva	1709	1978
Negativa	2858	2370
Negativa Positiva	1721	1853
Negativa Negativa	2860	2265

forma, foram escolhidos os dicionários: *crude*, dicionário do R, e *Stanford Political Dictionary* [6]. Após completar as palavras, foi dado prosseguimento ao processo de tradução automática através do Microsoft Translator.

Em que pese a tradução automática tenha sido efetiva para boa parte das palavras, foi preciso realizar o processo manualmente para alguns termos usando o Google Tradutor. Feito esse processo, a tradução do dicionário para o português do Brasil foi concluída para dar continuidade à Análise de Sentimento dos textos dos discursos, os quais estão em língua portuguesa do Brasil.

Como informação, o dicionário LSD contém palavras categorizadas em quatro polaridades distintas: positiva, negativa, negativa negativa e negativa positiva. Essa categorização foi mantida após a tradução do dicionário. Segue abaixo a definição de cada polaridade do dicionário:

- **Termo Positivo:** palavras que possuem uma conotação positiva. Por exemplo, *Hoje o dia está* ***lindo*** (palavra positiva **em negrito**);
- **Termo Negativo:** palavras que possuem uma conotação negativa. Por exemplo, *Hoje o dia está* ***trágico*** (palavra negativa **em negrito**);
- **Termo Negativo Positivo:** palavras que possuem uma conotação positiva, porém estão sendo negadas através do termo **não** e, com isso, se tornam negativas. Por exemplo, *Ele teve uma fase de* ***não aceitação*** *de seu problema* (palavra positiva e termo *não* **em negrito**);
- **Termo Negativo Negativo:** palavras que possuem uma conotação negativa, porém estão sendo negadas através do termo **não** e, com isso, se tornam positivas. Por exemplo, *Fique feliz,* ***não triste*** (palavra negativa e termo *não* **em negrito**).

Na Tabela 3.4, são apresentadas mais informações referentes ao dicionário LSD e ao dicionário criado. Cabe destacar que houve uma diferença no número de palavras entre os dois dicionários. Isso se deve ao fato de que algumas palavras em inglês podem ter vários sinônimos em português do Brasil e vice-versa. São características de cada língua. Por exemplo, *perdão* pode ser traduzido para o inglês como *forgiveness* e *pardon*.

Tabela 3.5: Exemplo da checagem das polaridades das palavras através da função *dfm* usando o Dicionário LSD traduzido.

ID do Texto/Polaridade	Negativa	Positiva	Neg. Positiva	Neg. Negativa
342555	156	216	0	0
410003	73	111	0	0
447973	127	196	0	0

3.6 Processamento dos discursos para realizar a análise de sentimento usando o dicionário traduzido

Após realizar a tradução do dicionário, foi dado o início à Análise de Sentimento dos discursos de senadores. À princípio, foram coletados aleatoriamente 18.000 discursos do Senado Federal entre 1980 e 2019. Em seguida, os textos dos discursos foram *tokenizados* e limpos, a fim de remover as *stopwords*, pontuações, caracteres especiais, espaçamento exagerado e números. Após, utilizou-se a função *dfm* do pacote do Quanteda, com o dicionário traduzido como parâmetro, para analisar as palavras dos discursos, com a finalidade de ver quantas palavras eram negativas, positivas, negativas positivas e negativas negativas no texto. Segue um exemplo de resultado da função *dfm*: Tabela 3.5. Cabe destacar que não foram encontradas palavras negativas negativas e negativas positivas. Com isso, essa abordagem não foi muito efetiva. Esse ponto será discutido mais a frente.

Depois de descobrir o número de palavras positivas, negativas, negativas positivas e negativas negativas para cada texto, calculou-se o *score* referente ao sentimento do texto. Sendo assim, o *score* está definido na Equação 3.1, o qual busca normalizar os valores encontrados, com a finalidade de realizar as análises e conferir o peso entre os termos positivos e negativos no texto. Essa equação foi usada com base no artigo [14] e no *cookbook* [16], o qual utiliza abordagem similar para calcular o *score* e a quantidade de palavras positivas, negativas, negativas positivas e negativas negativas. Para conhecimento, o *Numero(<polaridade>)* é a quantidade de palavras de certa polaridade.

$$ScoredeSentimento_{LSD} = log(\frac{Numero(Positiva) + Numero(NegativaNegativa) + 0,5}{Numero(Negativa) + Numero(NegativaPositiva) + 0,5}) \tag{3.1}$$

Cabe ressaltar que se o *score* tem o valor negativo, o texto tem mais termos negativos e, como consequência, apresenta uma opinião mais negativa quanto ao tema abordado. Se o *score* tem o valor positivo, ocorre o contrário, há mais termos positivos e, com isso, apresenta uma opinião mais positiva quanto ao tema.

Tabela 3.6: Exemplo da checagem das polaridades das palavras no texto - Dicionário OpLexicon v2.1.

Texto/Polaridade	**Positiva**	**Negativa**
342555	4	14
410003	10	5
447973	8	8

3.7 Processamento dos discursos para realizar a análise de sentimento usando outros dicionários

Para fazer um contraponto com o dicionário criado, o experimento foi feito também para os dicionários OpLexicon[4] versões 2.1 e 3.0 e SentiLex[5], encontrados nos trabalhos analisados no decorrer do projeto, como o [9], [8], [15], [12], entre outros.

Antes de mais nada, foi aproveitada a limpeza dos discursos (remoção das *stopwords*, pontuações, caracteres especiais, espaçamento exagerado e números) realizado no processamento do LSD anteriormente. Mesmo porque, os discursos escolhidos são os mesmos para fazer a comparação deles ao final. Em seguida, foi produzido um experimento usando um ferramental capaz de usar esses dicionários e fazer o processamento dos textos.

Nessa ceara, foram escolhidas as bibliotecas:

- **tm:** pacote que possui métodos para importação de dados, manipulação de corpus, construções de matrizes, entre outras funções;
- **stringr:** pacotes que permite manipular *strings* com mais facilidade;
- **tidytext:** pacote usado para fazer mineração de textos, por exemplo, foi usada a função *unnest_tokens* para tokenizar o texto;
- **dplyr:** pacote usado para manipulação de dataframes, por exemplo, o *inner_join* foi usado para checar a polaridade das palavras;
- **lexiconPT:** pacote que permite o uso de recursos léxicos da língua portuguesa.

Para realizar a análise léxica dos sentimentos, os discursos foram coletados dos arquivos *pickle*, transformados em *dataframe* e *tokenizados* para que a polaridade das palavras fossem checadas e pudessem ser classificada em *positivas* e *negativas*, quando for o caso. Na Tabela 3.6, na Tabela 3.7 e na Tabela 3.8, tem-se exemplos de textos e a quantidade de palavras *positivas* e *negativas* desses textos. Percebe-se que o LSD consegue classificar mais palavras em positivo e negativo que os demais dicionários.

Após a análise de sentimento, foi feito o cálculo de *score de sentimento* dos três dicionários, de forma bem similar e com o mesmo princípio do *score* do dicionário do LSD, conforme Equação 3.1. A Equação 3.2, a Equação 3.3 e a Equação 3.4 então foram montadas

Tabela 3.7: Exemplo da checagem das polaridades das palavras no texto - Dicionário OpLexicon v3.0.

Texto/Polaridade	**Positiva**	**Negativa**
342555	1	15
410003	6	6
447973	7	9

Tabela 3.8: Exemplo da checagem das polaridades das palavras no texto - Dicionário sentiLex.

Texto/Polaridade	**Positiva**	**Negativa**
342555	4	14
410003	10	6
447973	6	9

para medir o sentimento dos três dicionários. Cabe destacar que *Numero(<polaridade>)* é a quantidade de palavras de certa polaridade.

$$ScoredeSentimento_{OpLexiconv.2.1} = log(\frac{Numero(Positiva) + 0,5}{Numero(Negativa) + 0,5}) \tag{3.2}$$

$$ScoredeSentimento_{OpLexiconv.3.0} = log(\frac{Numero(Positiva) + 0,5}{Numero(Negativa) + 0,5}) \tag{3.3}$$

$$ScoredeSentimento_{sentiLex} = log(\frac{Numero(Positiva) + 0,5}{Numero(Negativa) + 0,5}) \tag{3.4}$$

3.8 Avaliação

Para fazer a avaliação do trabalho, depois de fazer o processamento dos dicionários, usou-se o método da inspeção, por não haver um método melhor. Mesmo porque, não há dado rotulado quanto ao sentimento dos discursos. Ao total, foram analisados 141 discursos e eles foram classificados em positivo e negativo manualmente, a partir de sua leitura e analisando o teor do texto. Foi escolhido esse número, pois foram selecionados os 20 maiores (com valores mais positivos) e os 20 menores (com valores mais negativos) s*cores de sentimento* para cada um dos quatro dicionários para realizar uma análise manual do sentimento do discurso. Os scores com valor menor que zero foram classificados pelo algoritmo como discurso com sentimento negativo. Já os com valor maior que zero, foram classificados como discurso com sentimento positivo.

Dessa forma, os resultados obtidos foram comparados usando os dicionários com os resultados encontrados na avaliação manual, com a finalidade de identificar se a avaliação por dicionário foi eficaz e conseguiria classificar um discurso em positivo e negativo.

Obviamente, não se pode dizer que é uma avaliação sistemática e que tem força suficiente para comprovar a hipótese dessa pesquisa (*é possível classificar os discursos em positivos e negativos, mesmo eles possuindo tamanhos variados e, em muitos casos, sendo maiores que tweets*), mas é uma maneira rápida de ter certos indícios e dar um direcionamento para pesquisas sobre esse assunto, já que considerou uma amostra representativa no contexto estudado. No entanto, essa parte ainda pode ser avançada para trazer resultados ainda mais consistentes. O resultado será abordado na próxima seção.

Capítulo 4

RESULTADOS

4.1 Comparação entre os Dicionários

Inicialmente, serão analisadas as quantidades de termos em cada dicionário, a fim de ver a diferença entre os dicionários e observar se há alguma correlação entre a quantidade de palavras no dicionário e a acurácia na hora do dicionário classificar o texto (acurácia seria o quanto o dicionário consegue classificar da mesma forma que a avaliação manual).

Após conferir a Tabela 4.1, percebe-se que o dicionário com maior quantidade de termos positivos é o OpLexicon v. 3.0, seguido do OpLexicon v. 2.1, do LSD em português do Brasil e do sentiLex. Já o com mais termos negativo também é o OpLexicon v. 3.0, seguido do OpLexicon v. 2.1, do sentiLex e do LSD em português do Brasil.

4.2 Resultado do Processamento dos Scores de Sentimento

Após realizar o cálculo do *score* para apresentar uma medida de sentimento do texto, os dados foram exportados e tabelados, com a finalidade de realizar uma conferência dos resultados pelo método da inspeção. Sendo assim, segue a Tabela 4.2 com exemplos de discursos e seus *scores de sentimento* para cada dicionário.

Cabe ressaltar que o discurso *350892* da Tabela 4.2 tem o seguinte trecho: *[...] Sr. Presidente, Sr^{a}s. e Srs. Senadores, o Governo Lula ainda tropeça nas contradições entre o que pregava no passado e o que agora faz. Mesmo assim, Lula renova promessas paralisadas até a metade de seu mandato, como é o caso do prometido aumento do investimento em educação. Artigo intitulado "A carroça do governo", publicado no Jornal do Brasil de 1^o de dezembro do corrente, mostra que o governo está atolado nas negociações com o PMDB para definir o indispensável apoio. Segundo o artigo, o governo perde um articu-*

Tabela 4.1: Quantidade de termos em cada dicionário usando neste trabalho.

Dicionário	**Polaridade**	**Quantidade dos Termos**
LSD em inglês	Positiva	1709
LSD em inglês	Negativa	2858
LSD em inglês	Negativa Positiva	1721
LSD em inglês	Negativa Negativa	2860
LSD em português do Brasil	Positiva	1978
LSD em português do Brasil	Negativa	2370
LSD em português do Brasil	Negativa Positiva	1853
LSD em português do Brasil	Negativa Negativa	2265
OpLexicon v2.1	Positiva	8524
OpLexicon v2.1	Negativa	14150
OpLexicon v3.0	Positiva	8620
OpLexicon v3.0	Negativa	14569
sentiLex	Positiva	1548
sentiLex	Negativa	4602

Tabela 4.2: Exemplo de discursos e seus *scores* de sentimento.

ID	**Score LSD**	**Score OpLexicon v3.0**	**Score OpLexicon v2.1**	**Score sentiLex**
350892	-0.363	-0.510	-0.510	-0.510
366679	1.058	1.098	1.098	1.098

lador influente, o deputado Jader Barbalho, denunciado por irregularidades no Banco do Estado do Pará. Assim, será difícil recuperar o tempo perdido, pois ao anunciar a reforma ministerial Lula teria dado a desculpa perfeita aos incompetentes para permanecerem inertes, dissimulando a ansiedade. Neste embaraçado cenário, a população ainda aguarda o cumprimento das promessas. Para que conste dos Anais do Senado, requeiro, Senhor Presidente, que o artigo citado, seja considerado como parte deste pronunciamento. [...]

E o discurso *366679* da Tabela 4.2 tem o seguinte trecho: *[...] Sr. Presidente, estou encaminhando à Mesa, em meu nome e em nome do Senador Sibá Machado, um requerimento de voto de aplauso para três cientistas brasileiros - os Professores Miguel Nicolelis, Cláudio Melo e Sidarta Ribeiro - e para todos os que apoiaram o Projeto do Instituto de Neurociência de Natal, que foi inaugurado semana passada. Esse instituto é extremamente importante por tratar-se de neurociência de ponta e, portanto, de uma área de educação científica extremamente relevante para o País. Inclusive, tive a oportunidade de ler uma extensa entrevista com o Professor Miguel Nicolelis publicada na Carta Capital há poucas semanas. Realmente, o trabalho que esse Instituto de Neurociência da Cidade de Natal vai desempenhar no País é muito importante. Então, estou encaminhando esse requerimento. Apenas gostaria de deixar o registro. [...]*

4.3 Análise do Resultado do Processamento dos Scores de Sentimento

Para analisar os scores de sentimento, foi realizado um histograma de frequência dos scores de sentimento para cada dicionário. Assim, foi possível ver a distribuição e o resultado e, com isso, fazer uma comparação da frequência dos scores entre os dicionários. O histograma do Dicionário LSD em português do Brasil é observado na Tabela 4.3, bem como o histograma do Dicionário OpLexicon v2.1, o histograma do Dicionário OpLexicon v3.0 e o histograma do Dicionário sentiLex.

Analisando os histogramas, percebeu-se que o LSD é o mais destoante, se comparado com os demais, já que a frequência de scores é maior no intervalo 0 a 0,5, chegando a uma frequência maior que 10.000. Os demais apresentam uma frequência de aproximadamente 3.000 e 4.000 scores nessa mesma faixa. Além disso, o histograma do OpLexicon v3.0 mostrou os scores mais dispersos, se comparado com os outros. A semelhança entre os quatro é que os discursos positivos são os mais frequentes.

4.4 Avaliação dos Resultados apresentados pelos Discursos Escolhidos para Inspeção

Conforme mencionado, o trabalho foi avaliado usando o método da inspeção, já que havia algumas limitações: dados não rotulados, um esforço muito grande rotular mais textos (foram avaliados 141 discursos) e falta de especialistas para avaliar os discursos. Mesmo assim, os resultados analisados podem apresentar, mesmo que preliminarmente, algum indício de que essa abordagem pode ser efetiva e ser um potencial de novos trabalhos que o amadureçam.

Diante do exposto, foi feita uma análise das palavras dos discursos escolhidos para participar da inspeção. Para isso, foi criada a Nuvem de Palavras da Figura 4.1, onde se percebeu que se compararmos com a outra Nuvem de Palavras (vide Figura 3.1), há a presença de muitas das palavras em ambas. Assim, nessa nova nuvem de palavras, as palavras mais frequentes foram: país, hoje, casa, nacional, agora, fazer, grande, partido, entre outras.

De modo geral, pegando o conceito de acurácia visto na seção 4.1, a acurácia dos dicionários foi significativa e é expressa na Tabela 4.4. De acordo com essa tabela, o sentiLex foi o dicionário que mais acertou a classificação do sentimento (positivo ou negativo) em 72,34% dos casos, tendo como base a classificação da Análise Manual. Por outro lado, o LSD Traduzido teve o pior resultado e acertou 57,45% dos casos.

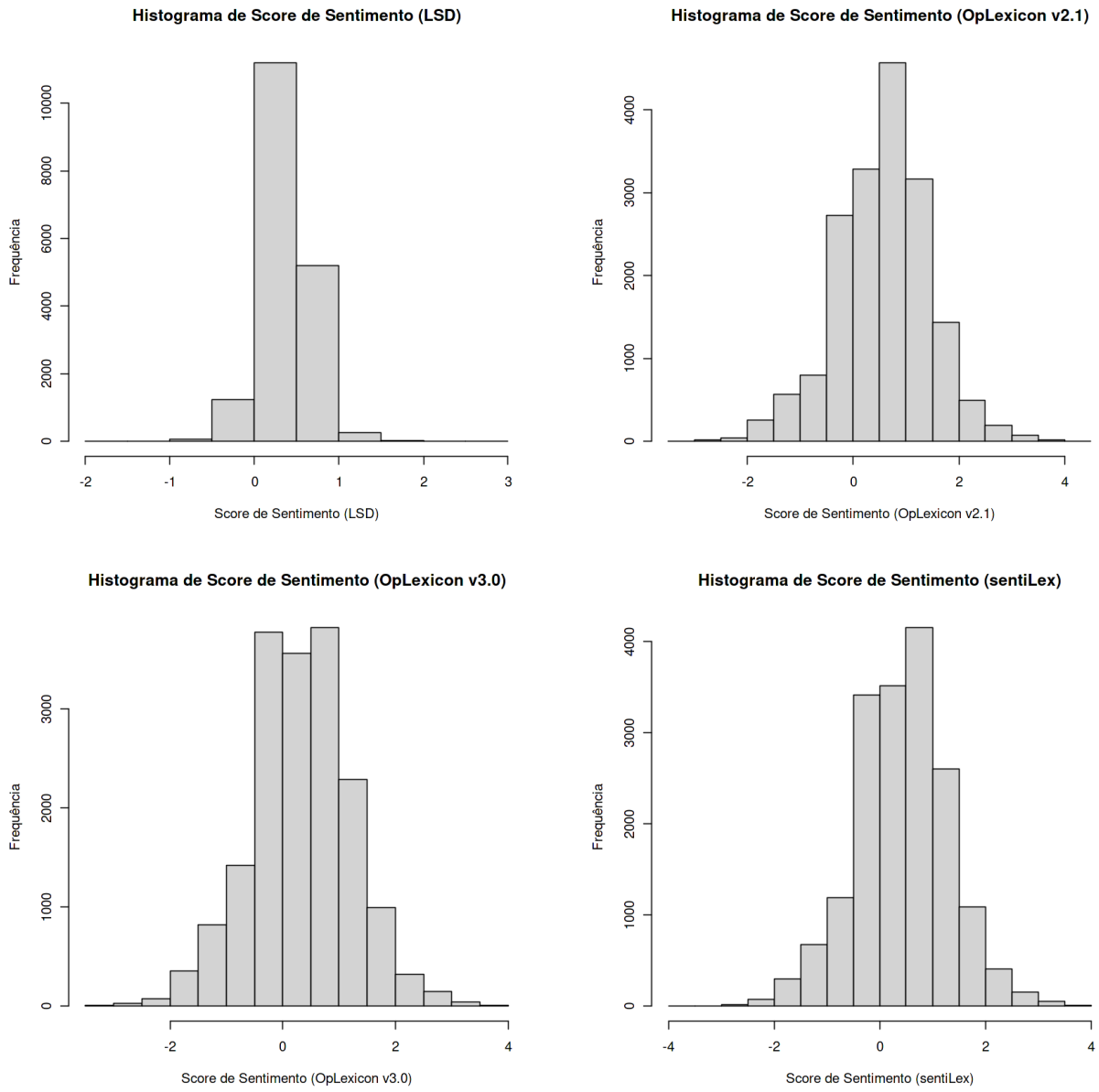


Tabela 4.3: Histograma dos Scores de Sentimento do Dicionários Analisados

Tabela 4.4: Resultado por Dicionário - Acurácia da Classificação de Sentimento se Comparada com a Análise Manual.

Dicionário	**Acurácia**
OpLexicon v2.1	64,54%
OpLexicon v3.0	67,38%
sentiLex	72,34%
LSD Traduzido	57,45%

Figura 4.1: Nuvem com as palavras mais frequentes dos discursos escolhidos para participar da inspeção

Agora, fazendo uma outra análise, a partir dos dicionários, foram coletadas as frequências dos discursos com sentimento positivo ou negativo por dicionário e análise manual. A Tabela 4.5 apresenta a frequência de polaridade positiva e a Tabela 4.6, a frequência de polaridade negativa. Cabe destacar, que os discursos que os dicionários OpLexicon v3.0 e v2.1 e sentiLex não conseguiram classificar, não foram listados nessas tabelas.

Analisando a Tabela 4.5, percebe-se que o LSD foi o dicionário que mais classificou os discursos como positivo (111 discursos) e o OpLexicon v3.0 o que menos classificou (68 discursos). Cabe mencionar que foram avaliados manualmente 51 discursos como positivo. O que mais chegou próximo a esse número foi o OpLexicon v3.0. Fazendo a média de frequência de discursos positivos nos dicionários, temos o valor de: 82,25 discursos.

Analisando a Tabela 4.6, percebe-se que o sentiLex e o OpLexicon v3.0 foram os dicionários que mais classificaram os discursos como negativo (63 discursos) e o LSD o que menos classificou (30 discursos). Cabe mencionar que foram avaliados manualmente 90 discursos como negativo. Os que mais chegaram próximo a esse número foram o sentiLex e o OpLexicon v3.0. Fazendo a média de frequência de discursos negativos nos dicionários, temos o valor de: 52 discursos.

Para entender melhor a proporção de positivos e negativos na avaliação manual e na avaliação dos dicionários, foram gerados gráficos, onde se pode ver a quantidade de discursos positivo e negativos. Em alguns casos, alguns discursos ficaram sem classificação. Nesses casos, os gráficos apresentaram o número de discursos nessa situação. O que

Tabela 4.5: Frequência dos discursos com sentimento positivo separados por dicionário ou análise.

Polaridade	Positivo				
Análise ou Dicionário	Manual	OpLexicon v3.0	OpLexicon v2.1	LSD	sentiLex
Frequência	51	68	80	111	70

Tabela 4.6: Frequência dos discursos com sentimento negativo separados por dicionário ou análise.

Polaridade	Negativo				
Análise ou Dicionário	Manual	OpLexicon v3.0	OpLexicon v2.1	LSD	sentiLex
Frequência	90	63	52	30	63

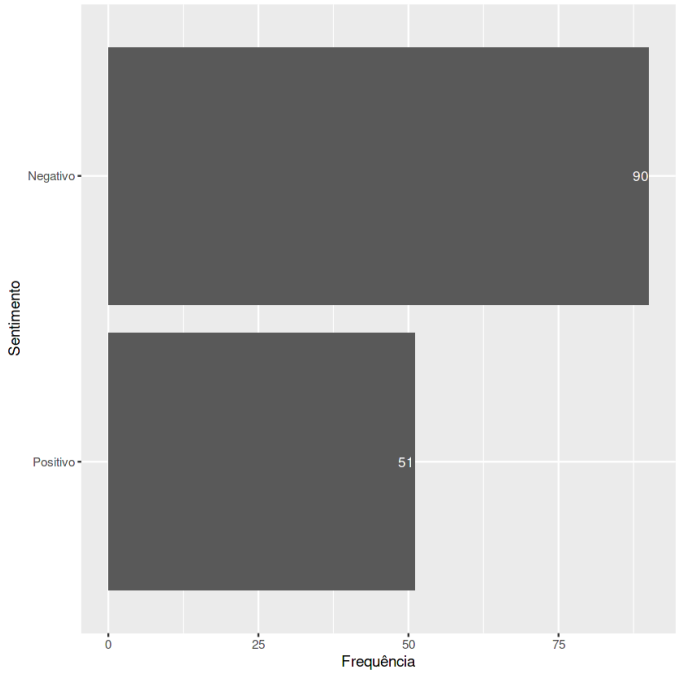
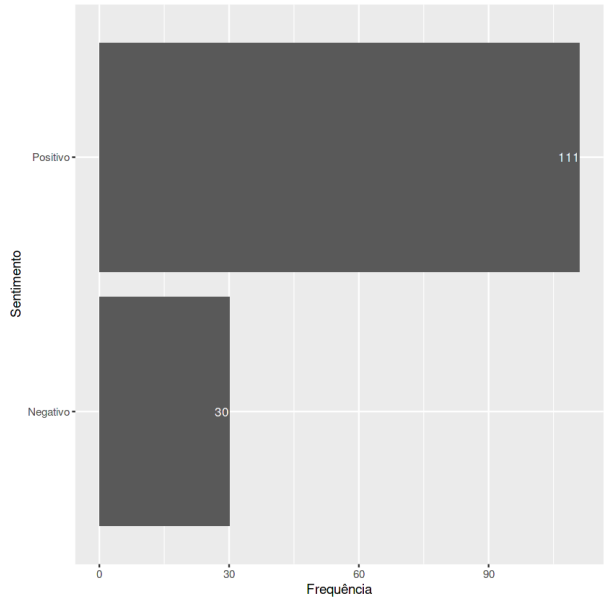


Figura 4.2: Gráfico com a Frequência de cada Polaridade na Avaliação Manual

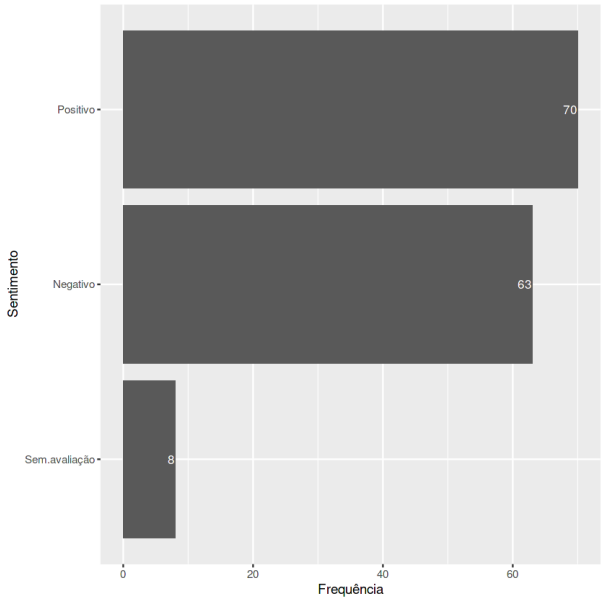
apresentou maior número de discursos sem avaliação foi o OpLexicon v3.0 (10 discursos). O gráfico da Avaliação Manual é observado na Figura 4.2 e os gráficos dos dicionários estão presentes na Tabela 4.7.

Para ter uma noção da quantidade de acertos dos dicionários, além da acurácia apresentada acima, foram avaliados os verdadeiros positivos, os verdadeiros negativos, os falsos positivos e os falsos negativos, tendo como base a análise manual para dizer se a classificação foi verdadeira ou falsa. A Tabela 4.8, a Tabela 4.9, a Tabela 4.10 e a Tabela 4.11 mostram as matrizes de confusão de cada dicionário. Analisando as matrizes, foram observados os seguintes pontos:

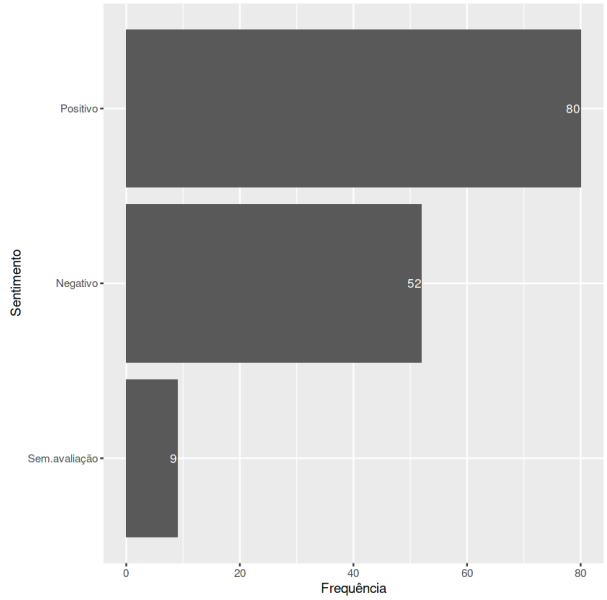
- Quanto ao dicionário LSD Traduzido, observou-se que não houve falso negativo. No entanto, houve muitos casos de falso positivo (60 casos);
- No geral, houve menos casos de falso negativo entre os dicionários, se comparado aos falso positivos;
- Os dicionários que apresentaram menos casos de falso negativo foram o sentiLex e o OpLexicon v2.1 (5 casos);
- O dicionário que apresentou mais casos de falso negativo foi o OpLexicon v3.0 (8 casos);
- Os dicionários que apresentaram menos casos de falso negativo foram o sentiLex e o OpLexicon v2.1 (5 casos);
- O dicionário que apresentou menos casos de falso positivo foi o sentiLex (26 casos).



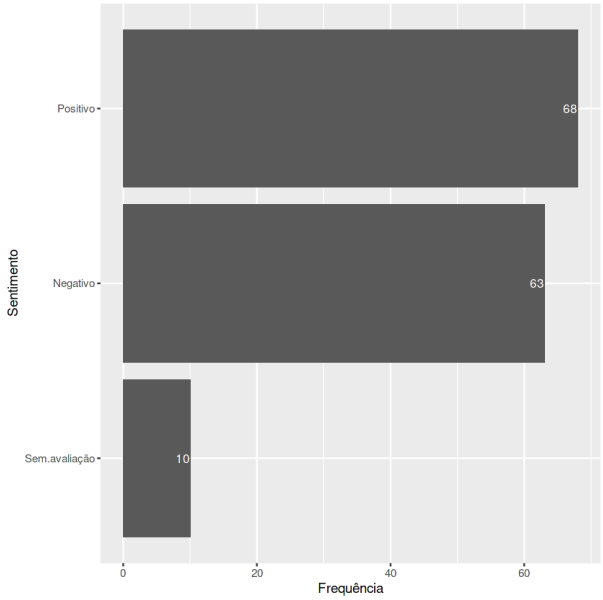
LSD Traduzido



SentiLex



OpLexicon v2.1



OpLexicon v3.0

Tabela 4.7: Gráfico com a Frequência de cada Polaridade na Análise Manual e nos Dicionários

Tabela 4.8: Matriz de Confusão para o Dicionário OpLexicon v3.0.

Polaridade/Referência	**Verdadeiro**	**Falso**
Positivo	40	28
Negativo	55	8

Tabela 4.9: Matriz de Confusão para o Dicionário OpLexicon v2.1.

Polaridade/Referência	**Verdadeiro**	**Falso**
Positivo	44	36
Negativo	47	5

Tabela 4.10: Matriz de Confusão para o Dicionário sentiLex.

Polaridade/Referência	**Verdadeiro**	**Falso**
Positivo	44	26
Negativo	58	5

Tabela 4.11: Matriz de Confusão para o Dicionário LSD Traduzido.

Polaridade/Referência	**Verdadeiro**	**Falso**
Positivo	51	60
Negativo	30	0

Capítulo 5

CONCLUSÃO

Na construção deste trabalho, concluiu-se que há diversas abordagens para se fazer análise de sentimento e ainda há muito o que se explorar, especialmente para a língua portuguesa do Brasil. Além disso, foi escolhido um objeto de estudo, discursos do Senado Federal, a fim de aplicar a análise de sentimento neste contexto e verificar sua potencialidade.

Os resultados foram positivos de certa forma. Um dicionário conseguiu acertar mais de 70% dos casos inspecionados, conforme visto no Capítulo 4. Em contrapartida, o dicionário construído (LSD Traduzido) conseguiu acertar, na avaliação dos discursos inspecionados, mais da metade dos casos e não houve falsos negativos. No entanto, houve muitos casos de falsos positivos. Esse ponto específico deve ser melhorado em projetos futuros com esse dicionário construído para que ele seja mais efetivo.

Cabe mencionar que a tradução do LSD foi a parte mais exaustiva e operacional do trabalho, dado ao volume de palavras para traduzir e conferir. Construir um dicionários realmente é um desafio e exige muita atenção. Recursos tecnológicos, como o Microsoft Translator, o Google Tradutor e os recursos da linguagem R, aceleraram a execução desse processo.

Além disso, durante a execução do trabalho, viu-se um ponto que exige atenção quanto ao dicionário LSD: ele é o único dentre os quatro estudados neste trabalho que possuem a classificação negativo negativo e negativo positivo. A ideia é muito interessante, tendo em vista que não necessariamente só ter palavras positivas ou negativas são o suficiente para dizer se um texto é positivo ou negativo. Às vezes, uma palavra positiva pode estar sendo negada com o uso do termo *não*, como visto no capítulo 2.

Dessa forma, seria uma boa ideia tentar entender como essa negação funciona nos textos para que a abordagem do dicionário seja efetiva na hora de identificar situações de negativo negativo e negativo positivo. Com os recursos usados neste experimento, cabe destacar que não houve efetividade na tentativa de encontrar esses casos. Neste contexto, seria valoroso melhorar a estratégia e as técnicas usadas para encontrar esses casos.

Quanto aos outros dicionários, o sentiLex foi o que teve a maior acurácia, mesmo apresentando alguns casos de falso positivo e falso negativo e discursos que não puderam ser classificados por esse dicionário. OpLexicon nas duas versões teve resultados similares. A versão 3.0 foi ligeiramente mais eficaz, quando se analisa os casos de falso positivo e falso negativo e acurácia. Dos três dicionários, o OpLexicon v2.1 teve a pior performance, porém foi na maioria dos pontos superior ao LSD Traduzido.

Um ponto visto durante o trabalho é o fato de ser fundamental a presença de especialistas da área dos dados estudados para orientar possíveis análises que são de interesse das pessoas dessa área. Os dados, no caso, são os textos de discursos de senadores e senadoras e a área é a Ciência Política. Neste trabalho específico, utilizaram-se os pontos explicitados dos pesquisadores da área de política do artigo [14]. Mesmo porque, de modo geral, avaliar o sentimentos não é uma tarefa simples pelo motivo de ser necessário saber um pouco mais sobre o contexto para poder analisar e classificar o sentimento dos textos, tendo em vista que cada área tem abordagens e visões, as quais podem ser distintas do senso comum para indicar o que é positivo e negativo.

A análise da frequência das palavras e o estudo dos meta-dados dos discursos foram essenciais para a construção deste trabalho, justamente por conta do ponto visto acima: a importância de se entender o contexto. É muito interessante como se pode usar recursos computacionais para gerar visualizações que facilitam a compreensão de algum dado. Neste trabalho, foram usados os recursos da linguagem R para renderizar os gráficos de barra, as tabelas e as nuvens de palavras. Esses recursos facilitaram muito a compreensão do contexto e trouxeram informações sobre os discursos. Os pacotes Quanteda [7] e tm foram muito relevantes para a construção do projeto, bem como outros recursos da linguagem R, pois com eles foi possível processar os dados e obter visualizações para fazer as análises.

A matriz de confusão foi um ótimo recurso para entender se houve falsos negativos e falsos positivos e auxiliou bastante o trabalho de avaliação dos dicionários quanto aos seus resultados em discursos escolhidos durante a inspeção. Em que pese o método de inspeção tenha suas limitações, ele era o único viável para a execução deste trabalho dada as limitações de recursos e pessoal especializado. Ele pode nos revelar resultados preliminares sobre a análise de sentimento neste contexto.

Quanto à hipótese de pesquisa: *é possível classificar os discursos em positivos e negativos, mesmo eles possuindo tamanhos variados e, em muitos casos, sendo maiores que tweets*, não se pode dizer com exatidão que é possível classificar os discursos em positivo e negativo neste contexto, mas os dicionários apresentaram resultados significativos, especialmente o sentiLex. Para responder essa pergunta com mais precisão, seria necessário realizar outros estudos similares e com tamanhos de discursos parecidos aos discursos

deste projeto.

De toda forma, um ponto a se notar é que a princípio não há uma relação clara entre o número de termos em um dicionário (esse ponto foi visto na Tabela 4.1) e a acurácia (esse ponto foi visto na Tabela 4.3). O artigo [14] inclusive abordou esse tema e chegou a conclusão que há um indício que esses pontos poderiam estar relacionados, sob a seguinte óptica: dicionários mais curtos e específicos (referentes a uma determinada área), quando aplicados na área, tendem a gerar resultados melhores. Neste trabalho de graduação, essa relação não pode ser comprovada. O pior resultado foi o LSD Traduzido e ele tem uma quantidade de termos mais parecida com o dicionário de melhor performance, sentiLex. Além disso, os resultados do sentiLex foram apenas ligeiramente melhores que os resultados das duas versões do OpLexicon, sendo que o OpLexicon tem muito mais termos que o sentiLex.

De modo geral, pode-se dizer que o trabalho apresentou resultados com um certo potencial. Entretanto, alguns pontos ainda devem ser melhorados ou explorados, como:

- melhorar a versão português do Brasil do dicionário LSD;
- melhorar as estratégias e técnicas de encontrar negativos negativos e negativos positivos;
- explorar outros dicionários;
- entender o porquê de haver muitos casos de falsos positivos;
- buscar especialistas da área de política para subsidiar a construção de futuros dicionários;
- aprimorar as técnicas de análise de sentimento usadas neste trabalho;
- usar outras abordagens de análise de sentimento.

Referências Bibliográficas

[1] Lei n. 12.527, de 18 de novembro de 2011, lei de acesso a informação. http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm. Acessado em: 2022-05-07. 1

[2] Portal de dados abertos do senado federal. https://www12.senado.leg.br/dados-abertos. Acessado em: 2022-04-13. 1

[3] Site com mais informações sobre o onto.pt. http://ontopt.dei.uc.pt/. Acessado em: 2022-04-13. 9

[4] Site com mais informações sobre o oplexicon. https://www.inf.pucrs.br/linatural/wordpress/recursos-e-ferramentas/oplexicon/. Acessado em: 2022-04-13. 9, 21

[5] Site com mais informações sobre o sentilex-pt 02. http://b2find.eudat.eu/dataset/b6bd16c2-a8ab-598f-be41-1e7aeecd60d3. Acessado em: 2022-04-13. 9, 21

[6] Site do dicionário stanford political dictionary. https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/VHXN5F. Acessado em: 2022-04-27. 19

[7] Site do pacote r quanteda para análise de dados como texto. https://quanteda.io/. Acessado em: 2022-04-27. 13, 16, 18, 34

[8] Pedro Balage Filho, Thiago Alexandre Salgueiro Pardo, and Sandra Aluísio. An evaluation of the brazilian portuguese liwc dictionary for sentiment analysis. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*, 2013. 2, 3, 21

[9] Fabrício Benevenuto, Filipe Ribeiro, and Matheus Araújo. Métodos para análise de sentimentos em mídias sociais. *Sociedade Brasileira de Computação*, 2015. 2, 9, 11, 21

[10] Edeleon Marcelo Nunes BRITO. Mineração de textos: detecção automática de sentimentos em comentários nas mídias sociais. *Universidade FUMEC, Belo Horizonte-MG*, 2017. 1, 2, 8, 11, 15

[11] Thales R Evangelista and Thereza P Pereira Padilha. Monitoramento de posts sobre empresas de e-commerce em redes sociais utilizando análise de sentimentos. In *Anais do III Brazilian Workshop on Social Network Analysis and Mining*, pages 152–163. SBC, 2014. 5

[12] Kássio TC Junqueira and Anita Maria da Rocha Fernandes. Análise de sentimento em redes sociais no idioma português com base em mensagens do twitter. *Anais do Computer on the Beach*, pages 681–690, 2018. 8, 9, 10, 21

[13] Anderson Pezzini. Mineração de textos: conceito, processo e aplicações. *Revista Eletrônica do Alto Vale do Itajaí*, 5(8):58–61, 2017. 5, 10

[14] Sven-Oliver Proksch, Will Lowe, Jens Wäckerle, and Stuart Soroka. Multilingual sentiment analysis: A new approach to measuring conflict in legislative speeches. *Legislative Studies Quarterly*, 44(1):97–131, 2019. 1, 2, 10, 12, 13, 14, 16, 18, 20, 34, 35

[15] Rômulo César Costa de Sousa. Identificando sentimentos de texto em português com o sentiwordnet traduzido. 2016. 2, 6, 7, 8, 9, 11, 21

[16] Lori Young. Lexicoder sentiment dictionary - codebook. 2015. 20

Anexo I

Tabelas

Tabela I.1: Frequência dos Discursos Analisados neste Trabalho por Partido.

Partido	**Frequência**
MDB	4161
PSDB	3793
PT	3657
PFL	1956
PDT	880
PSB	753
PTB	661
PP	514
PR	476
PDS	161
Sem Partido	150
PPB	138
PRN	116
PMB	108
PCdoB	104
PSD	81
PATRIOTA	63
PODEMOS	60
REDE	47
PPS	32
PSOL	31
PMN	11
PSL	11
DC	7
PODE	5
PTRB	4
PMR	3
PPR	3
PSC	3
DEM	2
PL	2
PTC	2
CIDADANIA	1
PPL	1
PRB	1
PROS	1
PV	1

Tabela I.2: Frequência dos Discursos Analisados neste Trabalho por Ano.

Ano	**Frequência**
1980	1
1994	104
1995	527
1996	719
1997	616
1998	404
1999	544
2000	538
2001	683
2002	396
2003	933
2004	885
2005	1242
2006	1232
2007	1029
2008	1076
2009	1081
2010	658
2011	1067
2012	643
2013	863
2014	389
2015	538
2016	539
2017	452
2018	373
2019	468