Universidade Federal de Ouro Preto
Instituto de Ciências Exatas e Biológicas
Departamento de Estatística
Bacharelado em Estatística

Análise de Sentimento dos Usuários do Twitter em Relação à Atual Situação Política do Brasil

Carolina Barcelos Teixeira

Ouro Preto-MG
Junho de 2019

Carolina Barcelos Teixeira

Análise de Sentimento dos Usuários do Twitter em Relação à Atual Situação Política do Brasil

Monografia de Graduação apresentada ao Departamento de Estatística do Instituto de Ciências Exatas e Biológicas da Universidade Federal de Ouro Preto como requisito parcial para a obtenção do grau de bacharel em Estatística.

Orientador

Prof. Dr. Tiago Martins Pereira

Universidade Federal de Ouro Preto – UFOP
Departamento de Estatística – DEEST

Ouro Preto-MG

Junho de 2019

T266a Teixeira, Carolina Barcelos.
Análise de sentimento dos usuários do twitter em relação à atual situação política do Brasil [manuscrito] / Carolina Barcelos Teixeira. - 2019.

59f.: il.: color; grafs; tabs; mapas.

Orientador: Prof. Dr. Tiago Martins Pereira.

Monografia (Graduação). Universidade Federal de Ouro Preto. Instituto de Ciências Exatas e Biológicas. Departamento de Estatística.

1. Tecnologia da informação. 2. Mineração de dados (Computação). 3. Emoções. I. Pereira, Tiago Martins. II. Universidade Federal de Ouro Preto. III. Titulo.

CDU: 004.77:316.64

Catalogação: ficha.sisbin@ufop.edu.br

Monografia de Graduação sob o título *Análise de Sentimento dos Usuários do Twitter em Relação à Atual Situação Política do Brasil* apresentada por Carolina Barcelos Teixeira e aceita pelo Departamento de Estatística do Instituto de Ciências Exatas e Biológicas da Universidade Federal de Ouro Preto, sendo aprovada por todos os membros da banca examinadora abaixo especificada:

Dr. Tiago Martins Pereira
Orientador
Departamento de Estatística
Universidade Federal de Ouro Preto

Dra. Diana Campos de Oliveira
Departamento de Estatística
Universidade Federal de Ouro Preto

Dr. Spencer Barbosa da Silva
Departamento de Estatística
Universidade Federal de Ouro Preto

Ouro Preto-MG, vinte e oito de junho de 2019.

Agradecimentos

Agradeço primeiramente à Deus, por guiar os meus passos, principalmente nos momentos mais difíceis. Por me iluminar em todas as escolhas e me amparar diante de todas as dificuldades.

Agradeço e dedico todo este trabalho aos meus pais, Kátia e Paulo, que através de todo amor, apoio e dedicação, tornaram esse sonho uma realidade.

Ao Leandro , pelo companheirismo, pela paciência e por ter sido meu ombro nas horas em que eu mais precisei.

Ao meu avô e aos meus tios e primos, por sonharem esse sonho junto comigo.

Agradeço também, aos meus amigos, que tornaram essa caminhada mais leve e gratificante; sobretudo à Juliana, Débora e Natália, por toda cumplicidade ao longo do curso.

À Universidade Federal de Ouro Preto pelo ensino de qualidade, pela estrutura oferecida e pelos excelentes profissionais.

Ao professor Tiago Martins Pereira pela parceria e suporte durante todo este trabalho, por despertar em mim a curiosidade sobre a Mineração de Dados, por todo aprendizado a mim oferecido e pela ajuda para que eu pudesse finalizar mais essa etapa em minha vida.

E por fim, não poderia deixar de agradecer à Estatis Jr. pela experiência incrível de poder vivenciar o mercado de trabalho ainda na graduação e à toda equipe do Departamento de Estatística pelo excepcional apoio prestado.

Quando o mundo estiver unido na busca do conhecimento, e não mais lutando por dinheiro e poder, então nossa sociedade poderá enfim evoluir a um novo nível.

Thomas Jefferson

Análise de Sentimento dos Usuários do Twitter em Relação a Atual Situação Política do Brasil

Autora: Carolina Barcelos Teixeira
Orientador: Prof. Dr. Tiago Martins Pereira

Resumo

As evoluções na área da Tecnologia da Informação possibilitaram o armazenamento de variadas e extensas bases de dados de natureza comercial, administrativa, governamental e científica. Por outro lado, a análise de grandes quantidades de dados pelo homem é inviável sem o auxílio de ferramentas computacionais apropriadas. Na era globalizada em que vivemos, é cada vez maior a competitividade entre as empresas, sendo a informação e o conhecimento elementos fundamentais para se obter diferenciais mercadológicos frente à concorrência. Diante desse cenário, surge a área de Mineração de Dados cujo propósito é desenvolver e aplicar técnicas que permitam obter conhecimentos novos e úteis a partir de grandes bases de dados, sejam elas numéricas ou textuais. A finalidade deste trabalho foi pesquisar e aplicar essas técnicas a fim de minerar dados vindos do Twitter, para a identificação do sentimento das pessoas em relação à atual situação política do Brasil, que em geral foi positiva.

Palavras-chave: Tecnologia da Informação, Mineração de Dados, Sentimento.

Twitter Users' Sentiment Analysis Regarding the Current Political Situation in Brazil

Author: Carolina Barcelos Teixeira
Advisor: PhD. Tiago Martins Pereira

Abstract

The evolutions in the area of Information Technology allowed the storage of varied and extensive databases of commercial, administrative, governmental and scientific nature. On the other hand, the analysis of large amounts of data by man is not feasible without the aid of appropriate computational tools. In the globalized era in which we live, the competitiveness between companies is increasing, with information and knowledge being fundamental elements in order to obtain market differentials in relation to the competition. Given this scenario, the area of Data Mining arises whose purpose is to develop and apply techniques that allow to obtain new and useful knowledge from large databases, be they numerical or textual. The purpose of this work was to research and apply these techniques in order to mine data coming from Twitter, to identify people's feelings about the current political situation in Brazil, which was generally positive.

Keywords: Information Technology, Data Mining, Sentiment.

Lista de figuras

Lista de tabelas

Lista de abreviaturas e siglas

UFOP – Universidade Federal de Ouro Preto

DEEST – Departamento de Estatística

KDD – *Knowledge Discovery in Databases*

RH – Relações Humanas

PLN – Processamento de Linguagem Natural

Sumário

1 Introdução

Os constantes avanços na área de tecnologia da informação têm viabilizado o armazenamento de grandes e múltiplas bases de dados. Internet, redes sociais, ambientes virtuais de aprendizagem, sistemas de telecomunicações e sistemas de informação em geral são alguns exemplos de recursos que têm tornado possíveis a criação e o crescimento de inúmeras bases de dados de natureza administrativa, científica, educacional, governamental e social (GOLDSCHMIDT; BEZERRA; PASSOS, 2015).

Segundo Cardoso e Machado (2008), a velocidade de coleta de informações tornou-se muito maior do que a velocidade de processamento ou análise das mesmas, gerando um problema e uma contradição, pois as organizações, por possuírem uma grande quantidade de dados, possuem uma falsa sensação de que estão bem informadas; porém essas informações de nada servem se não forem analisadas de forma correta e em tempo hábil.

O valor dos dados armazenados pode ser mensurado através da capacidade de extrair conhecimento de mais alto nível através deles, ou seja, informação útil que sirva para apoio à tomada de decisão. Um bom banco de dados pode apresentar padrões ou tendências, que se descobertas, podem se tornar de grande valia na tarefa de tomada de decisão.

Neste contexto, alguns questionamentos surgem naturalmente: O que fazer com todos os dados armazenados? Como utilizar as informações contidas em um banco de dados para benefício das instituições? Como analisar e utilizar todo o volume de dados disponível?

Com a finalidade de tentar responder à estas questões, foi proposta, no final da década de 80 a Mineração de Dados, do inglês *Data Mining*. A Mineração de Dados é a etapa de análise do processo conhecido como KDD - *Knowledge Discovery in Databases*, sendo a sua tradução literal "Descoberta de Conhecimento em Bases de Dados"e pode ser interpretada como um conjunto de técnicas de extração de informações e processos para investigar relacionamentos entre variáveis a partir de uma grande quantidade de dados. Nesse sentido, o objetivo da Mineração de Dados é obter padrões, anomalias e regras possibilitando a conversão de dados geralmente não perceptíveis em instrumento

para tomada de decisão gerando conhecimento e produzindo modelos preditivos.

O uso de Mineração de Dados ainda permite, por exemplo, que um supermercado melhore a disposição de seus produtos nas prateleiras, através do padrão de consumo de seus clientes; uma companhia de marketing direcione o envio de mensagens promocionais, obtendo melhores retornos; uma empresa aérea possa diferenciar seus serviços oferecendo um atendimento personalizado; empresas planejem melhor a logística de distribuição dos seus produtos, prevendo picos nas vendas; agências de viagens possam aumentar o volume de vendas direcionando seus pacotes a clientes com determinado perfil, dentre várias outras tarefas.

Existe ainda uma extensão da Mineração de Dados, dita Mineração de Textos, que possibilita a identificação de padrões em bases textuais. É uma técnica que se tornou muito usual nos dias atuais, uma vez que existem diversos dados não númericos capazes de gerar inúmeras informações úteis. É fácil perceber, por exemplo, a grande quantidade existente desses tipos de dados vindo da Internet, de redes sociais, por exemplo, que são extremamente utilizadas pelas pessoas.

Dentro da Mineração de Textos, surge a Análise de Sentimentos, que como o próprio nome já diz, nada mais é do que analisar o sentimento das pessoas com base no que elas dizem. Isso pode ser válido para avaliar se a opinião das pessoas é positiva ou negativa em relação a determinado produto, através de seus *tweets* postados no Twitter (tipo de rede social), por exemplo.

Sendo assim, a Mineração de Dados caracteriza-se como uma atividade multidisciplinar, envolvendo principalmente as áreas de Banco de Dados, Inteligência Artificial e Estatística. Também deve envolver conhecimentos específicos sobre os dados a serem minerados e os objetivos da entidade interessada na mineração, o que pode caracterizar também como uma área da Administração (HAN; KAMBER, 2006).

Goldschmidt, Bezerra e Passos (2015) argumentam que a Estatística é considerada ancestral da área de KDD . Técnicas de reconhecimento de padrões e de análise exploratória de dados provenientes da Estatística são muito utilizadas em algoritmos de Mineração de Dados. Amostragem, pré-processamento, transformação de dados e avaliação de padrões extraídos são apenas alguns exemplos de métodos estatísticos que são aplicados durante o processo de descoberta do conhecimento. A Estatística fornece uma linguagem para a quantificação da incerteza resultante quando se tenta inferir padrões a partir de uma amostra.

2 Objetivos

Foram objetivos deste trabalho:

- Pesquisar sobre os principais conceitos e técnicas relacionadas ao processo de descoberta do conhecimento através de Mineração de Dados;
- Pesquisar e aplicar técnicas de Mineração de Textos e Análise de Sentimentos;
- Aplicar técnicas estatísticas e computacionais de Mineração de Dados para análise dos sentimentos expressos nos *tweets* minerados a respeito da atual situação política do Brasil, que é um assunto extremamente comentado pelas pessoas.

3 Revisão de Literatura

Os padrões observados nos mais diversos episódios diários é uma verdadeira fonte de conhecimento para o ser humano. É assim que nascem hipóteses e teorias, impulsionando o crescimento e o desenvolvimento da sociedade. Com o advento da era tecnológica, essa ideia ganhou um novo direcionamento para abranger também a observação de informações geradas em ambientes digitais, dando origem a um processo denominado de Mineração de Dados.

Carvalho, Escobar e Tsunoda (2014) afirmam que vários são os desafios para a utilização da Mineração de Dados, que se inicia com a preparação das bases de dados, seleção dos algoritmos que melhor se adequam ao problema proposto e por fim análise dos resultados, ou seja, dos padrões descobertos. Por isso mesmo, apesar de todo desenvolvimento já realizado, esta área continua sendo objeto de pesquisas por novas soluções que se aproximem do real interesse dos potenciais usuários. As razões pela constante pesquisa se devem ao fato de que, apesar da existência de várias experimentações quanto ao uso de Mineração de Dados, ainda é baixa a sua adoção nos processos de tomada de decisão diários (MARISCAL; MARBAN; FERNANDEZ, 2010), por conta da pouca familiaridade dos especialistas com a metodologia e do desconhecimento das vantagens (MEYFROIDT et al., 2009) ou, ainda, a exploração comparativa de técnicas de Mineração de Dados, deixando de lado etapas relativas à interpretação e avaliação de resultados (BLOMBERG et al., 2010).

3.1 KDD - Descoberta de Conhecimento em Base de Dados

O termo KDD - *Knowledge Discovery in Databases*, foi formalizado em 1989 em referência ao amplo conceito de procurar conhecimento a partir de base de dados. Segundo Frawley, Piatetsky-Shapiro e Matheus (1991), "KDD é um processo, de várias etapas, não trivial, interativo e iterativo, para identificação de padrões compreensíveis, válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados". De acordo com

Goldschmidt, Bezerra e Passos (2015), o termo iterativo sugere a possibilidade de repetições integrais ou parciais do processo de KDD e a expressão não trivial alerta para a complexidade normalmente presente na execução de processos de KDD. Já com relação a expressão padrão válido indica que o conhecimento deve ser verdadeiro e adequado ao contexto da aplicação de KDD e o termo padrão novo deve acrescentar novos conhecimentos aos existentes, para que todo esse processo gere conhecimento útil que pode ser aplicado de forma a proporcionar benefícios ao contexto de aplicação de KDD.

Para proporcionar um melhor entendimento do problema, Goldschmidt, Bezerra e Passos (2015) salientam as diferenças e a escala de importância entre dado, informação e conhecimento, conforme sugere a Figura 1.

Figura 1: Escala de importância entre Dado, Informação e Conhecimento.

Ainda segundo esses autores, os dados podem ser interpretados como itens elementares, captados e armazenados por recursos da Tecnologia da Informação. São cadeias de símbolos e não possuem semântica. As informações representam os dados processados, com significados e contextos bem definidos, enquanto que o conhecimento corresponde a um padrão ou conjunto de padrões, cuja formulação pode envolver e relacionar dados e informações.

A pirâmide do conhecimento, representada na Figura 1, demonstra que a quantidade de dados existentes em um sistema é muito grande. Já o montante de informação é reduzido devido a dados errôneos ou sem expressão. Menor ainda é a quantidade de conhecimento que pode ser extraído desta informação através de técnicas de descoberta de conhecimento.

O processo de KDD é composto por cinco fases: seleção de dados, pré-processamento, transformação, mineração e interpretação/avaliação, conforme podemos observar na Figura 2.

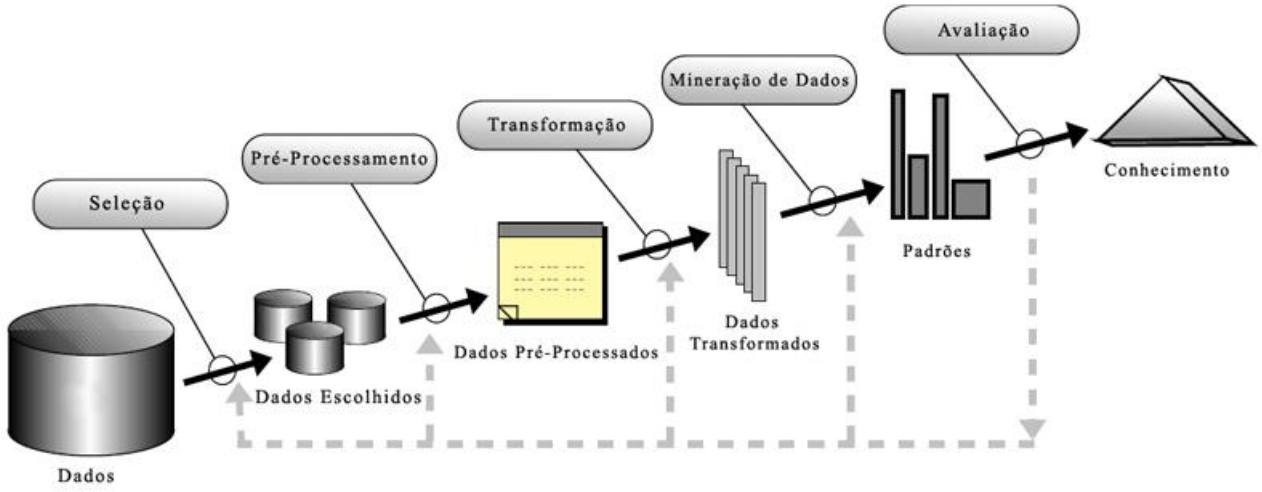

Figura 2: Representação do processo de KDD.

No decorrer deste processo pode-se encontrar conhecimentos explícitos ou não, isto é, as informações podem ser as que já se tinha conhecimento ou informações inesperadas que ao analisar a base de dados não se verificava nenhuma relação óbvia. Podem ainda ocorrer informações sem nenhuma relação significativa, pela falta de atributos ou por não haver conhecimento novo a ser descoberto.

Para um melhor entendimento de todo o processo, Goldschmidt, Bezerra e Passos (2015), sugerem que as fases do processo KDD sejam agrupadas em três grandes etapas denominadas por eles de etapas operacionais do processo KDD. A Figura 3 ilustra essas etapas.

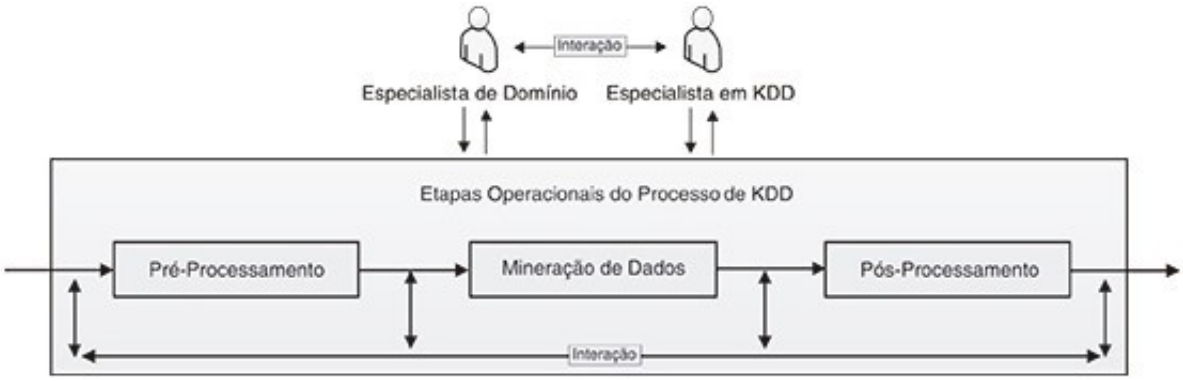

Figura 3: Etapas operacionais do processo de KDD.

3.1.1 Pré-Processamento

A etapa do pré-processamento compreende todas as funções relacionadas com a captação, organização e tratamento dos dados. Esta etapa tem como objetivo a preparação dos dados para os algoritmos de mineração de dados que serão aplicados na etapa seguinte. As principais funções de pré-processamento são destacadas abaixo:

- **Seleção de Dados:** Como já citado anteriormente, o processo KDD é a busca por conhecimento em grandes bases de dados. Sendo assim, torna-se inviável a aplicação dos algoritmos a todos eles, pois seria necessário o uso de máquinas muito potentes e demandaria muito tempo para a execução dos mesmos. Surge então a primeira etapa do processo, a seleção dos dados. Nessa fase, é feita a escolha dos dados que realmente serão utilizados no processo. A seleção dos dados pode ter dois enfoques distintos: a seleção de atributos ou a seleção de registros que devem ser submetidos ao processo de extração de conhecimento. Nesse estágio é interessante a participação de um especialista no conjunto de dados, uma vez que essa seleção é de fundamental impacto nos resultados finais das análises.

- **Limpeza dos Dados:** Em grandes bancos de dados é comum encontrar-se inconsistências: registros incompletos, valores incorretos e dados incoerentes. Assim sendo, essa etapa tem como finalidade eliminar essas adversidades de forma que elas não influenciem nos resultados finais. Algumas das causas desses problemas são erros humanos (erros de digitação, por exemplo), indisponibilidade da informação no momento do levantamento dos dados ou quando alguma mudança no sistema operacional ainda não tenha sido refletida no ambiente da Mineração de Dados. A retirada dos dados com problemas, a atribuição de valores padrões e a aplicação de técnicas de agrupamento, são algumas das técnicas utilizadas nessa etapa para auxiliar na descoberta dos melhores valores. É importante frisar que, a decisão de eliminar observações ou variáveis pode acarretar consequências nos resultados finais. Portanto, o ideal é o uso de técnicas para a substituição dos valores problemáticos, como pela média, mediana ou moda, por exemplo.

- **Integração dos Dados:** A função dessa etapa é analisar profundamente os dados coletados para que se possa integrá-los de forma consistente e apta à obtenção de resultados satisfatórios. Trata-se de uma fase extremamente importante, pois em grandes bases de dados é comum que nem todos tenham sido coletados da mesma forma e no mesmo local, o que pode comprometer a veracidade dos resultados ao fim

das análises. Redundâncias, dependências entre as variáveis e valores incompatíveis (categorias diferentes para os mesmos valores ou regras diferentes para os mesmos dados, por exemplo) são algumas técnicas observadas e corrigidas na integração dos dados.

- **Transformação dos Dados:** A transformação dos dados talvez seja a etapa mais importante do pré-processamento, pois é através dela que se constitui um padrão para os dados, facilitando o uso das técnicas computacionais de análises. Para isso, diversos métodos podem ser utilizados de acordo com os objetivos desejados. Na remoção de valores incorretos dos dados, por exemplo, pode-se utilizar a suavização, assim como se pode utilizar o agrupamento para agrupar os valores em faixas sumarizadas. Existe ainda a generalização que converte valores muito específicos em valores mais genéricos, a normalização que coloca as variáveis em uma mesma escala, a criação de novas características (gerados a partir de outras que já existem), entre outras.

- **Redução dos Dados:** Devido ao alto volume de dados estudados na mineração, em alguns casos, a redução dos dados é essencial para que análises sejam feitas de forma a se obter resultados satisfatórios, ainda que feita a seleção dos dados no início, pois quando os bancos são muito grandes, a mineração torna-se inviável. Assim, o objetivo dessa etapa é utilizar técnicas para que o conjunto de dados original seja convertido em um conjunto menor, porém sem perder o valor dos dados originais. Nessa etapa são aplicados mecanismos como a elaboração de estruturas otimizadas para os dados (cubos de dados), seleção de um subconjunto dos atributos, redução da dimensionalidade e discretização.

3.1.2 Mineração de Dados

Berry e Linoff (1997), definem a Mineração de Dados como a exploração e análise de grandes quantidades de dados, de maneira automática ou semiautomática, com o objetivo de descobrir padrões e regras relevantes. Para uma empresa, o conhecimento desses padrões permite melhorar a sua estratégia de marketing, personalizar seu atendimento, aprimorar seu relacionamento com os consumidores, etc. Para uma entidade pública, os conhecimentos adquiridos por uma Mineração de Dados podem subsidiar as suas ações políticas. Para o sistema financeiro, o conhecimento gerado pela base de dados de seus clientes permite obter critérios objetivos com vistas a discriminar e classificar bons e maus pagadores.

De acordo com Larose (2005), um dos fatores do sucesso da Mineração de Dados é o fato de dezenas, e muitas vezes centenas de milhões de reais serem gastos pelas companhias na coleta dos dados e, no entanto, nenhuma informação útil é identificada. Han e Kamber (2006) referem-se a essa situação como "rico em dados, pobre em informação". Além da iniciativa privada, o setor público e o terceiro setor (ONG's) também podem se beneficiar com a Mineração de Dados (WANG; HU; ZHU, 2008).

Segundo Written e Frank (2005), Olson e Delen (2008) e Bramer (2007), a Mineração de Dados vem sendo aplicada com sucesso em diversas áreas do conhecimento, encontrando padrões e gerando informações de forma satisfatória na retenção de clientes: identificação de perfis para determinados produtos, venda cruzada; em bancos: identificar padrões para auxiliar no gerenciamento de relacionamento com o cliente; em administradoras de cartão de crédito: identificar segmentos de mercado, identificar padrões de rotatividade; em agências de cobrança: detecção de fraudes; em empresas de telemarketing: acesso facilitado aos dados do cliente; em pesquisas eleitorais: identificação de um perfil para possíveis votantes; na medicina: indicação de diagnósticos mais precisos; em segurança: na detecção de atividades terroristas e criminais; no auxílio a pesquisas biométricas; em RH : identificação de competências em currículos; na tomada de decisão: filtrar as informações relevantes, fornecer indicadores de probabilidade, etc.

Sendo assim, esta etapa consiste no uso de técnicas que permitem automatizar a busca em grandes volumes de dados por padrões e tendências que não são detectáveis por análises mais simples, auxiliando gestores na tomada de decisões estratégicas. As tarefas mais comuns aplicadas na prática de Mineração de Dados incluem:

- **Sumarização:** descreve de forma compacta um subconjunto de dados. Essa função é frequentemente utilizada, por exemplo, na análise exploratória de dados, com geração automatizada de relatórios.
- **Classificação:** Uma das tarefas mais comuns, a classificação, visa identificar a qual classe um determinado registro pertence. Nesta tarefa, o modelo analisa o conjunto de registros fornecidos, com cada registro já contendo a indicação à qual classe pertence, a fim de "aprender"como classificar um novo registro (aprendizado supervisionado). Por exemplo, categorizamos cada registro de um conjunto de dados contendo as informações sobre os colaboradores de uma empresa: Perfil Técnico, Perfil Negocial e Perfil Gerencial. O modelo analisa os registros e então é capaz de dizer em qual categoria um novo colaborador se encaixa. A tarefa de classificação pode ser usada por exemplo para:

1. Determinar quando uma transação de cartão de crédito pode ser uma fraude;
2. Identificar em uma escola, qual a turma mais indicada para um determinado aluno;
3. Diagnosticar onde uma determinada doença pode estar presente;
4. Identificar quando uma pessoa pode ser uma ameaça para a segurança.

- **Estimação ou Regressão:** A estimação é similar à classificação, porém é usada quando o registro é identificado por um valor numérico e não um categórico. Assim, pode-se estimar o valor de uma determinada variável analisando-se os valores das demais. Por exemplo, um conjunto de registros contendo os valores mensais gastos por diversos tipos de consumidores e de acordo com os hábitos de cada um. Após ter analisado os dados, o modelo é capaz de dizer qual será o valor gasto por um novo consumidor. A tarefa de estimação pode ser usada por exemplo para:

  1. Estimar a quantia a ser gasta por uma família de quatro pessoas durante a volta às aulas;
  2. Estimar a pressão ideal de um paciente baseando-se na idade, sexo e massa corporal.

- **Predição:** A tarefa de predição é similar às tarefas de classificação e estimação, porém ela visa descobrir o valor futuro de um determinado atributo. Exemplos:

  1. Predizer o valor de uma ação três meses adiante;
  2. Predizer o percentual que será aumentado de tráfego na rede se a velocidade aumentar;
  3. Predizer o vencedor do campeonato baseando-se na comparação das estatísticas dos times.

- **Agrupamento ou Clusterização:** A tarefa de agrupamento visa identificar e aproximar os registros similares. Um agrupamento (ou *cluster*) é uma coleção de registros similares entre si, porém diferentes dos outros registros nos demais agrupamentos. Esta tarefa difere da classificação pois não necessita que os registros sejam previamente categorizados (aprendizado não-supervisionado). Além disso, ela não tem a pretensão de classificar, estimar ou predizer o valor de uma variável, ela apenas identifica os grupos de dados similares. As aplicações das tarefas de agrupamento são as mais variadas possíveis: pesquisa de mercado, reconhecimento de padrões,

processamento de imagens, análise de dados, segmentação de mercado, taxonomia de plantas e animais, pesquisas geográficas, classificação de documentos da *Web*, detecção de comportamentos atípicos (fraudes), entre outras. Geralmente a tarefa de agrupamento é combinada com outras tarefas, além de serem usadas na fase de preparação dos dados.

- **Descoberta de Associações:** A tarefa de associação consiste em identificar quais atributos estão relacionados. Apresentam a forma: se atributo X, então atributo Y. É uma das tarefas mais conhecidas devido aos bons resultados obtidos, principalmente nas análises da "Cestas de Compras"(*Market Basket*), onde identificamos quais produtos são levados juntos pelos consumidores. Alguns exemplos:

  1. Determinar os casos onde um novo medicamento pode apresentar efeitos colaterais;
  2. Identificar os usuários de planos que respondem bem a oferta de novos serviços.

- **Detecção de Desvios:** Esta tarefa consiste em identificar observações do conjunto de dados cujas características não atendam aos padrões considerados normais no contexto. Tais observações são denominadas de observações atípicas ou *outliers*.

- **Descoberta de Sequências:** É uma extensão da tarefa de descoberta de associações na qual são buscados itens frequentes levando-se em conta várias transações ocorridas ao longo de um período de tempo.

Deve-se ressaltar que cada técnica de Mineração de Dados ou cada implementação específica dos algoritmos que são utilizados para conduzir as operações dessas técnicas, adequa-se melhor a alguns problemas que a outros, o que dificulta a existência de um método inteiramente melhor. À vista disso, o êxito de uma tarefa de Mineração de Dados está diretamente vinculado à intuição e experiência do analista.

Feita essa etapa, é possível a realização das análises sobre o banco de dados e utilização dos conhecimentos adquiridos a cerca do mesmo.

3.1.3 Pós-Processamento

Segundo Goldschmidt, Bezerra e Passos (2015), a etapa do pós-processamento abrange o tratamento do conhecimento adquirido na etapa de mineração de dados. Entre as principais funções desta etapa destacam-se a elaboração e organização do conhecimento obtido,

podendo incluir a simplificação de gráficos, diagramas e relatórios, além da conversão da forma de representação em conhecimento obtido.

3.2 Mineração de Textos

A Mineração de Textos é uma extensão da Mineração de Dados, e pode ser definida como um processo de extração de informações desconhecidas e úteis de documentos textuais escritos em linguagem natural. Segundo Tan et al. (1999), 80% das informações de uma companhia estão contidas em documentos textuais. Chen (2001) afirma ainda que 80% do conteúdo online estão em formato texto. A partir destas afirmações, conclui-se que apenas 20% das informações são usadas para manipulação de tomada de decisão dentro das empresas.

A Mineração de Textos surge, então, da necessidade de se descobrir, de forma automática, informações (padrões e anomalias) em textos. Trata-se de um conjunto de métodos usados para navegar, organizar, achar e descobrir informações em bases textuais. Pode ser vista como uma extensão da área de *Data Mining*, focada na análise de textos. Segundo Passos e Aranha (2006), a Mineração de Textos é um campo multidisciplinar que inclui conhecimentos de áreas como Informática, Estatística, Linguística e Ciência Cognitiva.

Com base no conhecimento extraído dessas ciências, a Mineração de Textos define técnicas de extração de padrões ou tendências de grandes volumes de textos em linguagem natural, normalmente, para objetivos específicos. Inspirado pelo *Data Mining* ou Mineração de Dados, que procura descobrir padrões emergentes de banco de dados estruturados, a Mineração de Textos pretende extrair conhecimentos úteis de dados não estruturados ou semi-estruturados.

A descoberta de conhecimento em textos conta, basicamente, com duas etapas: a primeira etapa consiste no tratamento do texto a fim de convertê-lo para uma forma estruturada; a segunda etapa consiste na aplicação da mineração para a descoberta do conhecimento (CORRÊA et al., 2003)

As técnicas de Processamento de Linguagem Natural (PLN) e Recuperação de Informação são aplicadas na primeira etapa e a Descoberta de Conhecimento em Banco de Dados é aplicada na segunda etapa.

3.2.1 O Processo de Mineração de Textos

De forma geral, as etapas do processo de Mineração de Textos são as seguintes: coleta de documentos, preparação dos dados, indexação e normalização, mineração de dados e pós-processamento (análise de resultados). Estas etapas podem ser visualizadas através do diagrama de atividades representado na Figura 4.

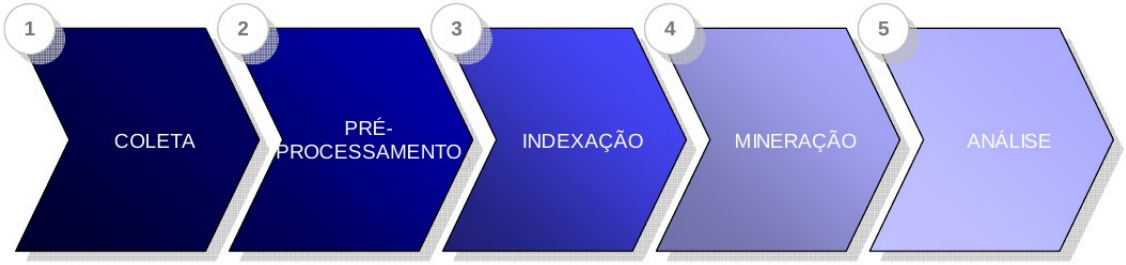

Figura 4: O processo de mineração de textos.

Coleta é a etapa inicial e tem como objetivo formar uma base de dados textual, conhecida na literatura como *Corpus* ou Corpora. Pode se dar de várias maneiras, porém todas necessitam de grande esforço, a fim de se conseguir material de qualidade e que sirva de matéria-prima para a aquisição de conhecimento.

Pré-processamento é a etapa executada imediatamente após a Coleta e tem como objetivo prover alguma formatação e representação da massa textual. É bastante onerosa, com a aplicação de diversos algoritmos que consomem boa parte do tempo do processo de extração de conhecimento. Para contribuir na etapa de pré-processamento de dados de textos, técnicas de PLN (Processamento de Linguagem Natural) podem ser utilizadas, como remoção de *stopwords* (palavras muito comuns que aparecem no texto e carregam pouco significado, como "as", "e", "os", etc.), segmentação de palavras, lematização, dentre outras.

Indexação é o processo que organiza todos os termos adquiridos a partir de fontes de dados, facilitando o seu acesso e recuperação. Uma boa estrutura de índices garante rapidez e agilidade ao processo, tal como funciona o índice de um livro.

Após terem sido obtidas uma estrutura para os dados e uma forma de prover rápido acesso, a etapa de mineração propriamente dita é responsável pelo desenvolvimento de cálculos, inferências e algoritmos e que tem como objetivo a extração de conhecimento, descoberta de padrões e comportamentos que possam surpreender.

Finalmente, a Análise é a última etapa e deve ser executada por pessoas que, nor-

malmente, estão interessadas no conhecimento extraído e que devem tomar algum tipo de decisão apoiada no processo de Mineração de Texto.

3.3 Análise de sentimentos em redes sociais

A análise de sentimentos é um campo de estudo com recente popularização devido ao crescimento da Internet e do conteúdo que é gerado por seus usuários, principalmente nas redes sociais, nas quais as pessoas publicam suas opiniões em uma linguagem coloquial e em muitos casos utilizando de artifícios gráficos para tornar ainda mais sucintos seus diálogos. A partir da explosão das redes sociais de uso global, a análise de sentimentos começou a ter um valor social muito importante. A Figura 5 endossa tal situação ao exibir o crescimento na quantidade de buscas pelo termo "*Sentiment Analysis*"no buscador Google à partir de 2008 *(TRENDS, 2019)*.

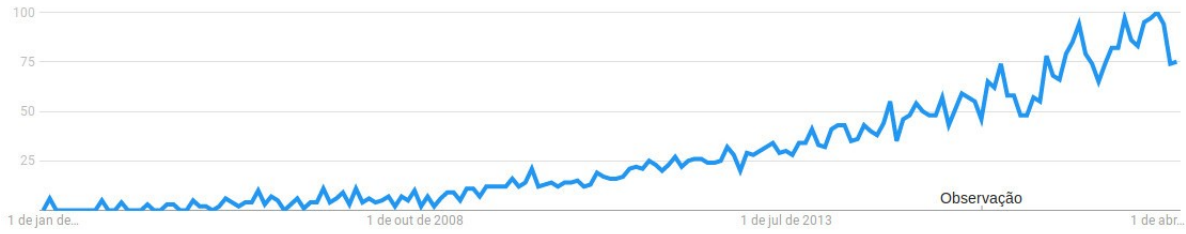

Figura 5: Pesquisa do termo "*Sentiment Analysis*"no *Google Trends*.

A facilidade para difusão de informações oferecidas pelas redes sociais e seu papel na sociedade moderna representam uma das novidades mais interessantes desses últimos anos, captando o interesse de pesquisadores, jornalistas, empresas e governos. A interligação densa que muitas vezes surge entre os usuários ativos gera um espaço de discussão que é capaz de motivar e envolver indivíduos ilustres ou influentes em discussões, ligando pessoas com objetivos comuns e facilitando diversas formas de ação coletiva. As redes sociais são, portanto, a criação de uma revolução digital, permitindo a expressão e difusão das emoções e opiniões através da rede. De fato, redes sociais são locais onde as pessoas discutem sobre tudo expressando opiniões políticas, religiosas ou mesmo sobre marcas, produtos e serviços.

Opiniões têm grande influência sobre o comportamento das pessoas. Decisões simples como qual carro comprar, qual filme ver, ou em qual ação investir são frequentemente baseadas em opiniões de pessoas próximas, de especialistas, ou de estudos conduzidos por

instituições especializadas. Organizações baseiam suas estratégias de negócio e investimentos na opinião de seus clientes sobre seus produtos ou serviços.

A expansão das mídias sociais disponibiliza a indivíduos e organizações conteúdo de opinião diversificado e em grandes volumes. Contudo, o grande volume de informação produzido diariamente implica a necessidade de métodos e ferramentas capazes de processar automaticamente não apenas o conteúdo das publicações, mas também a opinião e sentimento que expressam.

Segundo Silva (2016), o Twitter é atualmente o *microblog* mais popular na Internet e possui, como principal característica, permitir aos usuários divulgar o que estão fazendo em tempo real, para todos os usuários ligados à sua rede (seus seguidores) com um limite máximo de 140 caracteres. Essas micro-mensagens são denominadas de *tweets*. Com mais de 313 milhões de usuários ativos mensalmente, o Twitter continuamente sofre alterações para manter seus usuários ativos na rede, adaptando-se às suas necessidades. Uma das mais significativas nos últimos anos foi o aumento do número de caracteres de 140 para 280.

De acordo com Liu (2012), a mineração de opiniões opera sobre porções de texto de quaisquer tamanho e formato, tais como páginas *web*, *posts*, comentários, *tweets*, revisões de produto, etc. Toda opinião é composta de pelo menos dois elementos chave: um alvo e um sentimento sobre este alvo. Um alvo pode ser uma entidade, aspecto de uma entidade, ou tópico, representando um produto, pessoa, organização, marca, evento, etc. Já um sentimento representa uma atitude, opinião ou emoção que o autor da opinião tem a respeito do alvo. A polaridade de um sentimento corresponde a um ponto em alguma escala que representa a avaliação positiva, neutra ou negativa do significado deste sentimento (TSYTSARAU; PALPANAS, 2012).

3.3.1 Classificação de sentimentos

Segundo Hekima (2014), a forma mais comum de classificação de sentimentos é definida pela polaridade. As publicações são divididas entre negativas ou positivas. Quando a opinião do usuário não é explícita, a publicação é considerada neutra. A vantagem da análise polar é a objetividade. Com apenas três opções, a classificação é simplificada e exige menos tempo.

A classificação por escala tem como objetivo definir diferentes níveis para as publicações positivas e negativas. Através de uma escala de 1 a 5, por exemplo, um *post* pode ser

considerado muito ou pouco negativo/positivo. Esse tipo de análise pode ser eficiente para diferenciar críticas e comentários de grande impacto de publicações menos significativas. O problema dessa forma de classificação é que a definição de números em uma escala é bastante subjetiva. Por isso, é difícil manter um mesmo padrão de classificação entre diferentes analistas, e a consistência da análise pode ficar comprometida.

Outra forma de classificar publicações é através da análise de aspecto ou característica por elemento. Um usuário pode, ao comentar sobre um *smartphone*, por exemplo, expressar diferentes opiniões sobre diferentes aspectos do aparelho. Em uma mesma publicação, o *design* pode receber uma avaliação positiva, enquanto o sistema operacional é criticado. Nesse caso, a maneira mais básica de classificação negativo/positivo não é eficaz. A ideia da análise de aspecto ou característica é classificar cada elemento separadamente. Dessa forma, a publicação é dividida em núcleos, e cada elemento citado é analisado isoladamente.

Por fim, a análise de humor é uma maneira mais aprofundada de classificar publicações. O objetivo é ir além do "positivo"ou "negativo", e definir o estado de espírito do usuário. A classificação pode variar entre termos como "insatisfação", "frustração", "revolta", "satisfação", "alegria", "expectativa", entre outros.

3.3.2 Abordagem Léxica para Análise de Sentimentos

Léxico é o conjunto ou acervo de palavras que um determinado idioma possui. Portanto, a análise léxica estuda as unidades do vocabulário, ou seja, as palavras portadoras de sentido: substantivos, adjetivos, verbos, advérbios entre outras.

Outra maneira de se analisar um texto seria a análise sintática, que se encarrega de examinar, classificar e reconhecer as estruturas da sintaxe, isto é, os períodos, as orações e os termos das orações. E por fim, a análise de um texto também pode ser feita através da análise semântica, que verifica o seu significado.

Na análise léxica, a intensidade do sentimento pode ser substituída por uma polaridade positiva, negativa ou neutra, como exemplifica a Tabela 1, construída por Pang, Lee et al. (2008).

Tabela 1: Tabela de palavras positivas e negativas.

	Palavras
Estudante 1	positivas: brilhante, fenomenal, excelente, fantástico
	negativas: terrível, horroroso
Estudante 2	positivas: espetacular, legal, excelente
	negativas: ruim, estúpido, lerdo

Também é comum a construção de um dicionário de palavras, que neste caso, se diferencia dos dicionários convencionais, pois o significado das palavras é substituído pela classificação numérica, que indica o valor do sentimento.

Um dicionário léxico de sentimento é uma espécie de dicionário de palavras que ao invés de possuir como conteúdo o significado de cada palavra, possui em seu lugar um significado quantitativo (pode ser um número entre -1 a 1, onde -1 é o valor sentimental mais negativo e 1 o valor mais positivo) ou mesmo valor qualitativo (positivo/negativo, feliz/triste).

A análise de sentimentos baseadas em abordagens léxicas é atualmente uma das estratégias mais eficientes, seja na utilização de recursos computacionais, seja em capacidade de predição. Ela se baseia na utilização de um grande dicionário de termos, onde cada termo está associado a um sentimento. A Figura 6 generaliza o funcionamento de um método de análise de sentimentos léxico. O processo de classificação inicia quando o método recebe uma sentença de entrada, em seguida é realizado um processamento de linguagem natural, assim como uma pesquisa no léxico dos termos que formam esta mensagem. Ao final do processo o método é capaz de inferir qual é a polaridade ou sentimento implícito na sentença de entrada.

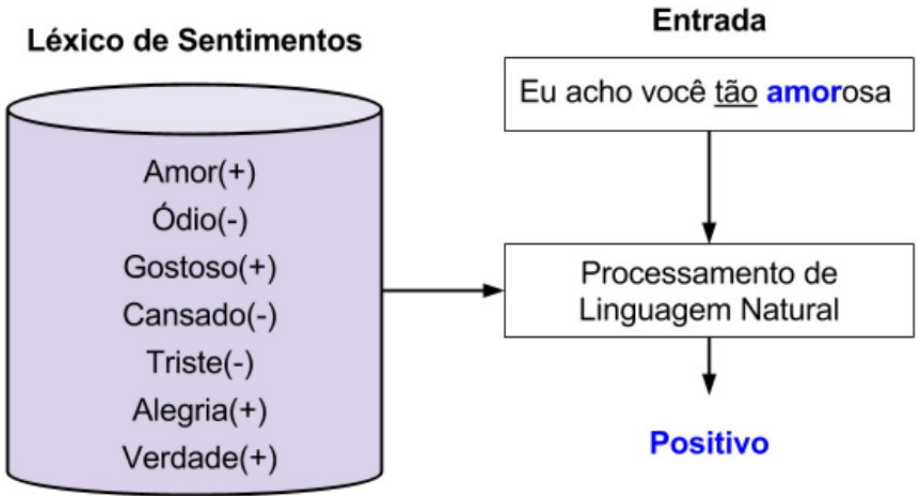

Figura 6: Léxico de sentimentos.

Por fim, palavras que não possuem valor agregador de sentimentos em um texto, tais como: são, meu, uma, os, a, dessa, por, para e aquilo, são denominadas *stopwords* (ARMANO; FANNI; GIULIANI, 2015). As *stopwords* não precisam ser consideradas na análise de um texto. Elas não tem um valor que incremente ou diminua a análise de sentimentos (DENECKE, 2008). Segundo Braga (2009) há benefícios na remoção de *stopwords* de uma frase para a análise sentimental. Para identificar quais são as palavras que não possuem valor agregador de sentimento e polaridade em uma frase, pode-se considerar uma lista existente de *stopwords* ou criá-la de acordo com as necessidades, tanto no caso dos dicionários como nas análises efetuadas por métodos de aprendizagem de máquina.

3.3.3 Nuvem de Palavras

Nuvens de palavras são imagens compostas por palavras de várias cores e tamanhos e, opcionalmente, organizadas em direções distintas, para demonstrar de maneira visual, a frequência da ocorrência das palavras em um dado texto, isto é, quanto maior for o número de ocorrências de uma palavra, maior a mesma será na nuvem de palavras.

A Nuvem de Palavras é uma importante ferramenta para a análise dos assuntos mais comentados pelos usuários, quando acessam o Twitter. Por ser um relatório muito visual, a Nuvem pode auxiliar no processo de análise da opinião do usuário devido à sua facilidade de compreensão.

4 Metodologia

Este trabalho foi estruturado em 3 etapas: (i) Obtenção dos dados, (ii) Pré-processamento dos *tweets*, (iii)Análise dos *tweets*.

4.1 Obtenção dos dados

A coleta dos dados do Twitter foi a primeira parte da execução deste trabalho. Para esta etapa, foi utilizado um *script* escrito na linguagem R (TEAM, 2018), que recebe como parâmetro palavras-chave e *hashtags* e retorna *tweets* que contenham as mesmas. Este processo de coleta foi no dia 24 de maio de 2019 e foram obtidos 107240 *tweets*, utilizando as palavras e *hashtag*s: reforma, previdência, educação, universidades e corte. Ao final desta fase, iniciou-se a etapa de pré-processamento.

4.2 Pré-processamento dos *tweets*

Nesta etapa, os *tweets* passaram por um pré-processamento, onde foram descartados *tweets* repetidos e conteúdos considerados irrelevantes para o processo de classificação de texto, como *links*, nomes de usuário (no Twitter, marcados pelo caractere "@") e caracteres não alfabéticos, com exceção do caractere "#", que é utilizado para marcar uma palavra como *hashtag*.

Ainda nesta fase foi aplicada a remoção de *stopwords*, uma técnica de Processamento de Linguagem Natural para a retirada de palavras que não possuem relevância para os resultados da classificação de textos. Estas foram descartadas dos conjuntos de dados, melhorando o desempenho do algoritmo de classificação de textos que seria utilizado posteriormente. Para a obtenção das *stopwords*, foi utilizado um conjunto de palavras disponibilizadas em NL (2018).

Para a construção dos dicionários léxicos utilizados na classificação dos *tweets* de

acordo com sua polaridade, positiva ou negativa, foi utilizada a metodologia proposta por Freitas (2013) e o dicionário léxico de sentimentos disponível em Tatman (2017).

Após feita toda a adequação necessária ao número total de *tweets* coletados, passou-se a ter 98106 para a realização das análises.

4.3 Análise dos *tweets*

A última etapa deste trabalho foi a análise dos tweets coletados e pré-processados. Através de gráficos foi possível identificar o sentimento das pessoas com relação ao assunto estudado e possíveis interpretações para esse resultado.

5 Resultados e Discussões

Foi gerada uma nuvem de palavras a fim de reproduzir graficamente as variações das palavras citadas pelas pessoas com mais frequência, onde as que mais se destacam aparecem em tamanhos maiores.

É possível notar na Figura 7 que a palavra "educação"possui a maior frequência, aparecendo 40470 vezes e se destacando em meio à tantas palavras mencionadas nos *tweets*.

As palavras "reforma"e "previdência"aparecem 30099 e 25328 vezes, respectivamente, seguidas das palavras "corte", "brasil", "governo", "ministro", "universidades"e "bolsonaro", que apareceram 12183, 7354, 7278, 5789, 5743 e 5445 vezes, de modo respectivo.

Observa-se ainda um destaque significante em palavras como "manifestação", "diabrasilnasruas", "fazer", "paulo", "guedes", "anticrime", "aprovação", "moro", "vamos", "presidente", "congresso", "povo", "nova", "economia", "saúde"e "tributária", o que pode apontar um possível sentimento positivo com relação à atual situação política brasileira, uma vez que a coleta de *tweets* foi realizada dois dias antes da Manifestação Pró Bolsonaro, ocorrida em 26 de maio de 2019 em vários locais no Brasil.

Figura 7: Nuvem de Palavras.

Para uma melhor visualização, foi gerada uma nuvem limitada às 200 palavras mais frequentes. Através da Figura 8, é possível observar de forma nítida quais são as palavras que mais se sobressaem dentre o total de *tweets*.

Em destaque obviamente aparece a palavra "educação", seguida das palavras "reforma"e "previdência". As palavras "corte", "governo", "brasil"e "bolsonaro"também aparecem com um destaque um pouco mais significante que as demais, assim como mostrado anteriormente.

A diferença é que nesta nuvem de palavras é possível enxergar de forma clara além destes termos evidenciados e verificar assertivamente quais as palavras aparecem mais vezes frente ao total de *tweets* analisados.

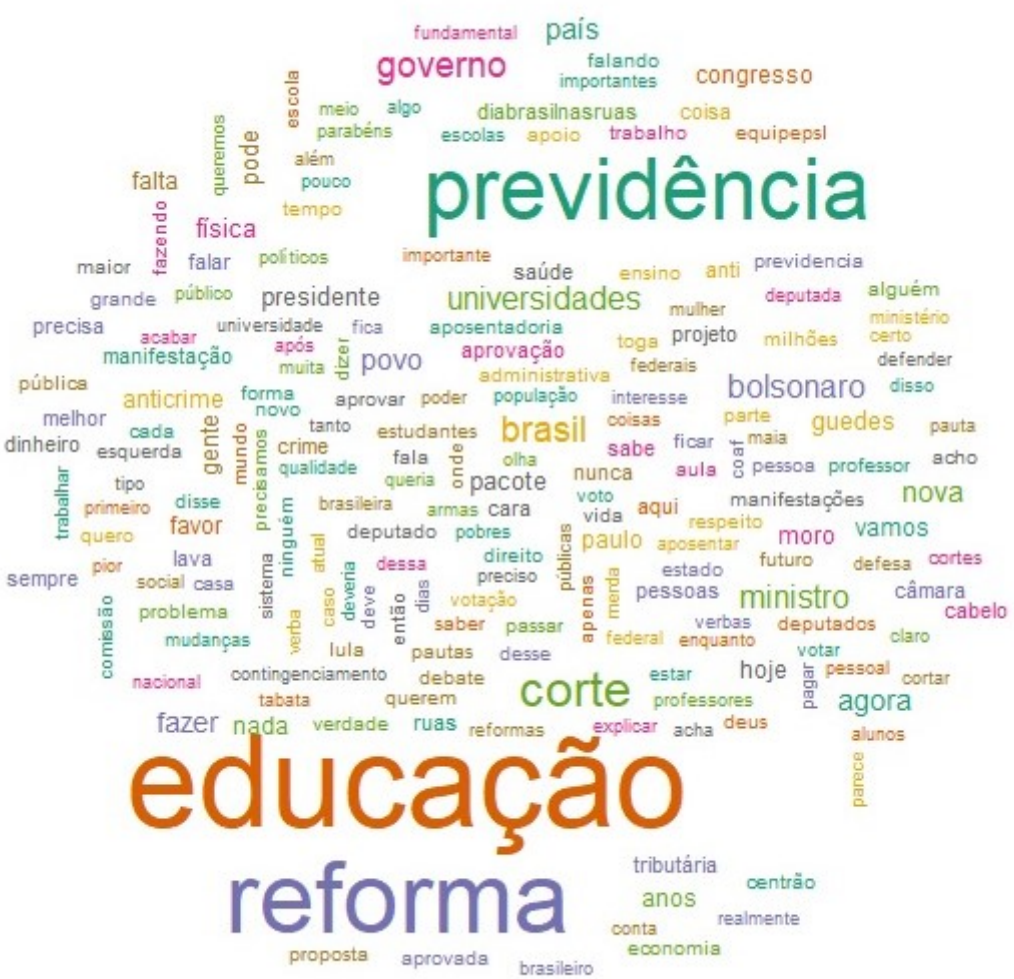

Figura 8: Nuvem das 200 palavras mais frequentes.

Em seguida foram calculadas as frequências de cada *hashtag* relacionadas aos *tweets* coletados, como mostrado na Tabela 2.

"Sem tag"se refere àqueles *tweets* que contêm às palavras em questão, porém que não possuem de fato as *hashtags*.

Tabela 2: Tabela de frequências das *hashtags*.

Hashtag	Frequência
educação	37966
reforma	26480
sem tag	15532
corte	11158
universidades	4688
previdência	2282

As Figuras 9 e 10 possibilitam verificar de forma gráfica a diferença de frequência entre as *hashtags* utilizadas, no qual educação e reforma aparecem em alta, cerca de 39% e 27%, respectivamente, seguidas dos *tweets* sem tag, 16% aproximadamente.

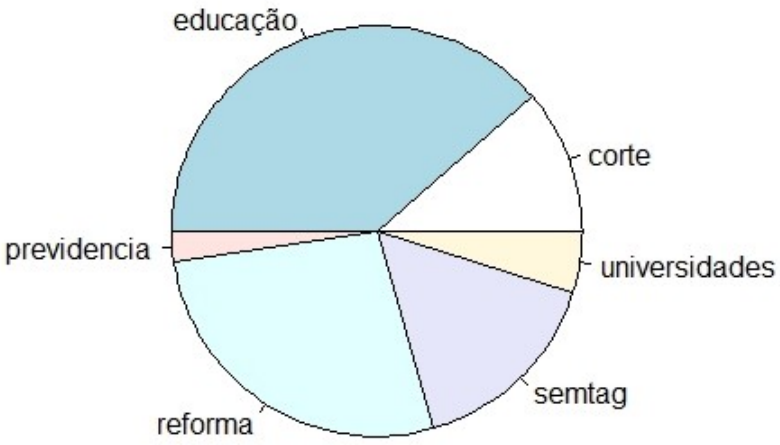

Figura 9: Gráfico de setores das frequências das *hashtags*.

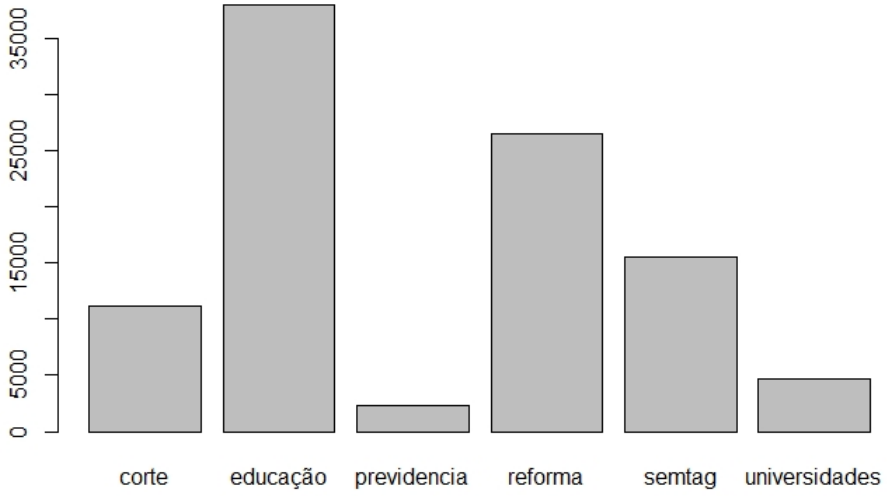

Figura 10: Gráfico de barras das frequências das *hashtags*.

Fica explícita uma baixa frequência da *hashtag* "previdência", o que pode causar

certo estranhamento uma vez que nas nuvens de palavras é uma palavra extremamente frequente. Isso acontece porque a palavra foi mencionada diversas vezes, mas não necessariamente a *hashtag* propriamente dita, o que pode explicar inclusive a alta quantidade de *tweets* sem tag.

Também foi possível identificar e representar graficamente a relação entre as *hashtags* mais utilizadas conjuntamente nos *tweets* analisados, como está exposto na Figura 11. Nota-se que as pessoas que utilizam a *hashtag* #dia26euvou, tem grandes chances de também utilizarem #bolsonaronossopresidente ou #reformadaprevidencia, #dia26brasilnasruas e #pacoteanticrimejá, por exemplo, o que também pode ser um indício de favoritismo ao atual governo por parte da maioria das pessoas.

Hashtags como #tsunaminaeducação, #impeachmentbolsonaro, #educacao, #lulalivre, #naoareformadaprevidencia e #forabolsonarodia30vaisermaior aparecem mais isoladas. Vale ressaltar que dia 30 de maio aconteceria a segunda manifestação contra o atual governo, mas fica evidente que a manifestação a favor, já mencionada anteriormente, estava sendo o assunto no momento.

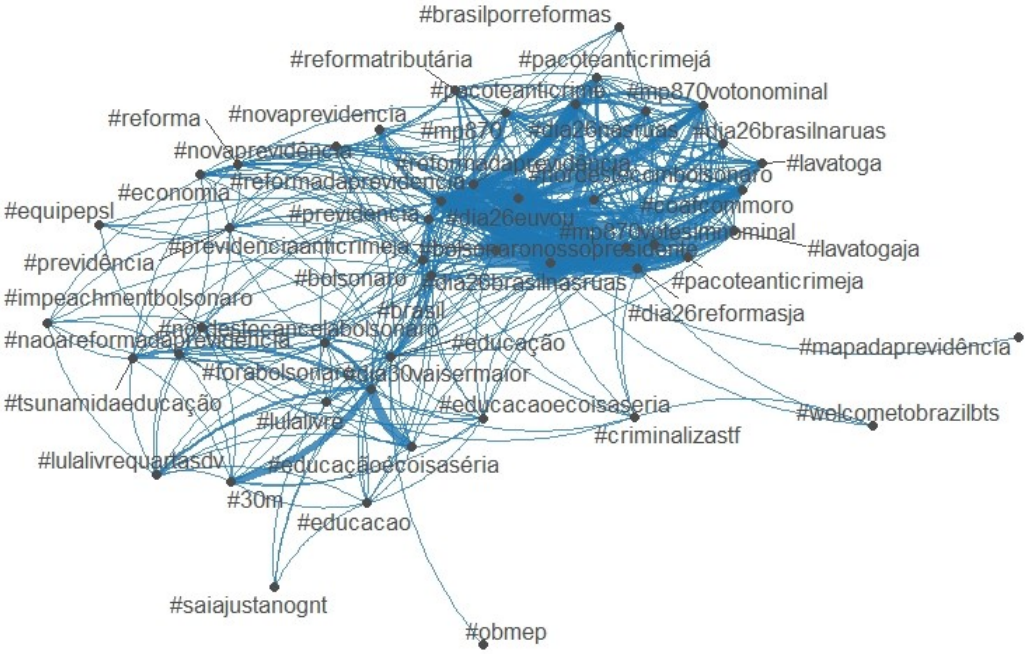

Figura 11: Gráfico da relação entre as *hashtags*.

O próximo passo foi analisar a frequência com relação ao total de *tweets*, ou seja, diante de todos, quais palavras mais apareceram. Na Tabela 3 observa-se os 20 termos que mais se sobressaem nos *tweets* de forma geral e as suas respectivas frequências, sendo possível observá-los graficamente na Figura 12.

Já na Figura 13 é possível notar a frequência referente à cada expressão, ou seja, à cada duas palavras que mais aparecem conjuntamente, das 25 mais comentadas pelos usuários do Twitter, estando evidenciados os valores relativos às 6 expressões mais em alta na Tabela 4.

Tabela 3: Tabela de frequências das palavras que mais se destacam.

Hashtag	Frequência
educação	40470
reforma	30099
previdência	25328
corte	12183
vai	9249
brasil	7354
governo	7278
ser	7092
dia	5975
ministro	5789
universidades	5743
bolsonaro	5445
contra	5264
sobre	5199
povo	4743
nova	4650
país	4266
vc	4221
fazer	4218
agora	4153

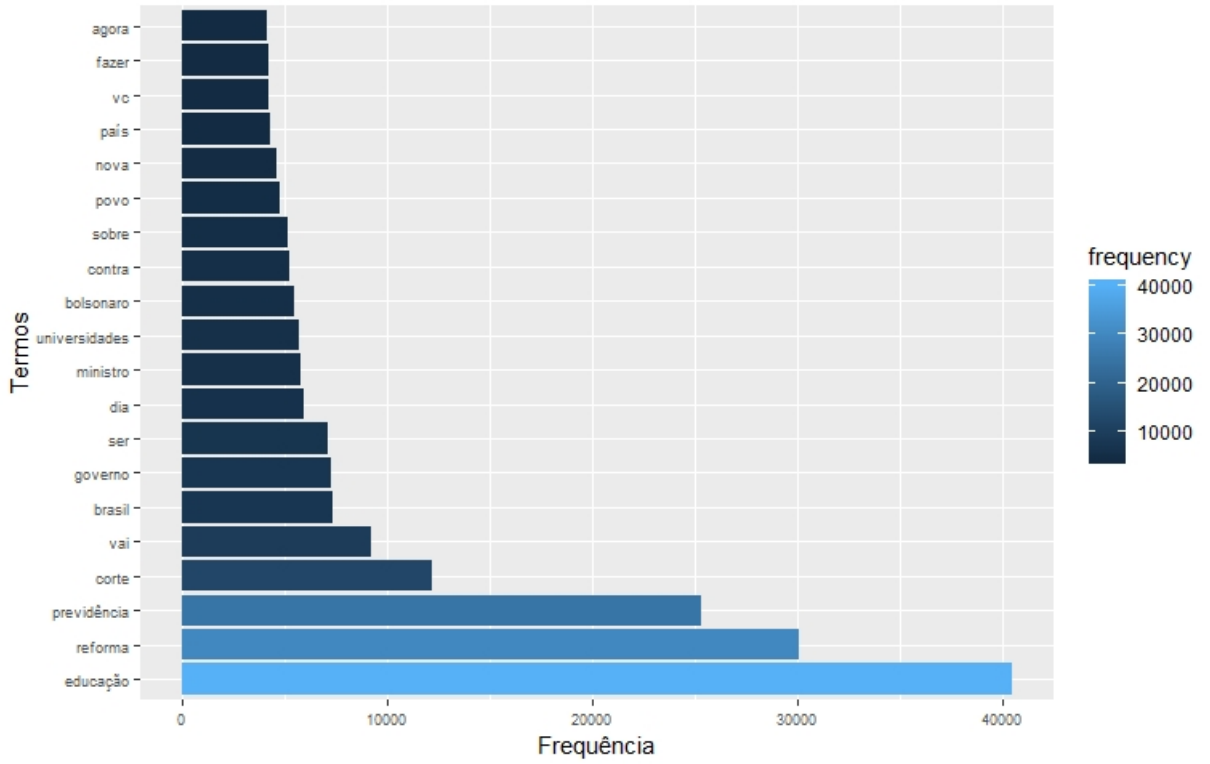

Figura 12: Gráfico de frequências das palavras que mais se destacam.

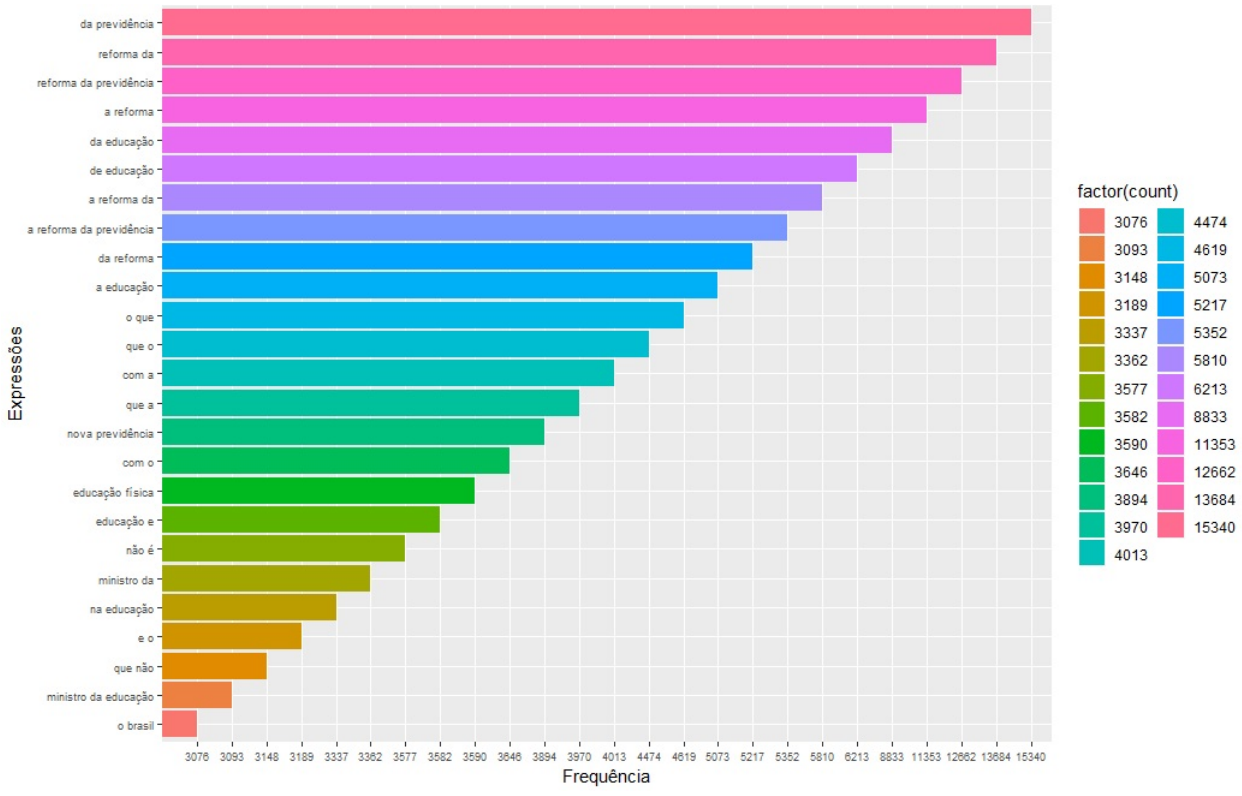

Figura 13: Gráfico de frequências das expressões que mais se destacam.

Tabela 4: Tabela de frequências das 6 expressões que mais se destacam.

Expressão	Frequência
da previdência	15340
reforma da	13684
a reforma	11353
da educação	8833
paulo guedes	2034
pacote anticrime	1816

E, finalmente, foi feita a análise de sentimentos das pessoas em relação à atual situação política do Brasil. Para isso, os dados foram estruturados e posteriormente analisados em relação ao dicionário léxico de sentimentos.

Através da Figura 14 fica fácil visualizar a porcentagem da frequência por polaridade, negativa ou positiva, de cada *hashtag* utilizada inicialmente, tornando possível a comparação e identificação de qual sentimento os *tweets* com cada *hashtag* se identificam mais.

Isso fica ainda mais evidente na Figura 15 que apresenta a mesma informação de forma mais clara, através de gráficos de barras, em que a cor rosa se refere ao sentimento negativo e a cor azul ao positivo.

Nota-se, através da Figura 15, que o sentimento da maioria das pessoas com relação ao corte de verba na educação é negativo, mas com relação à reforma da previdência, o sentimento que sobressai é o positivo.

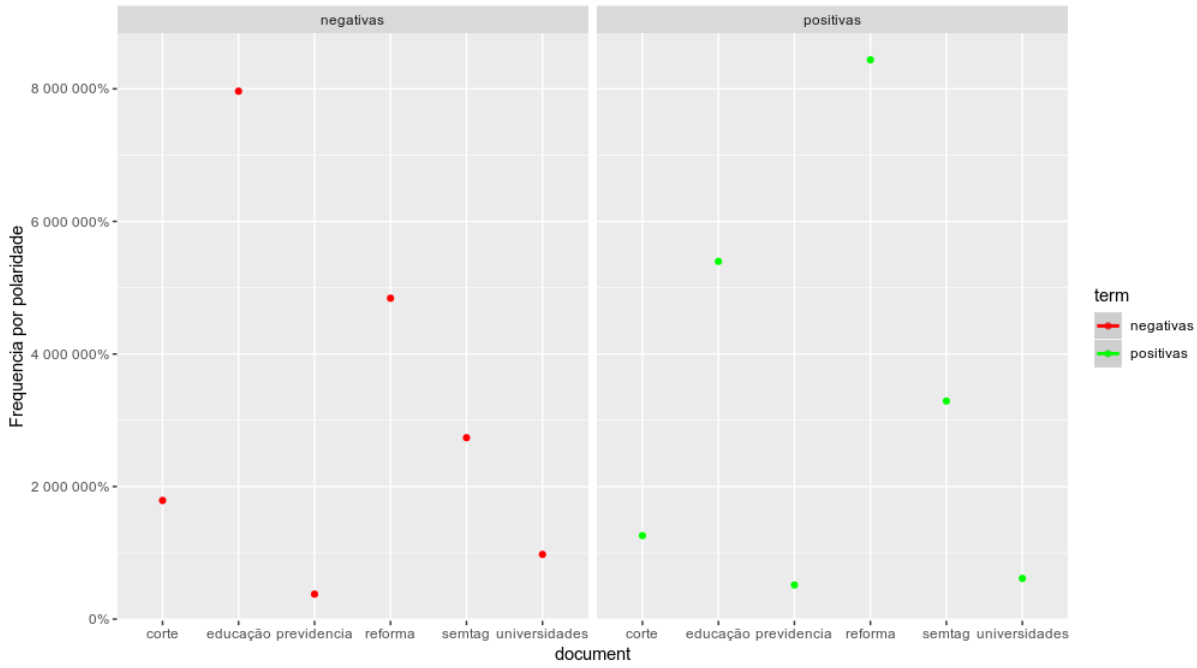

Figura 14: Gráfico da frequência relativa de cada *hashtag*.

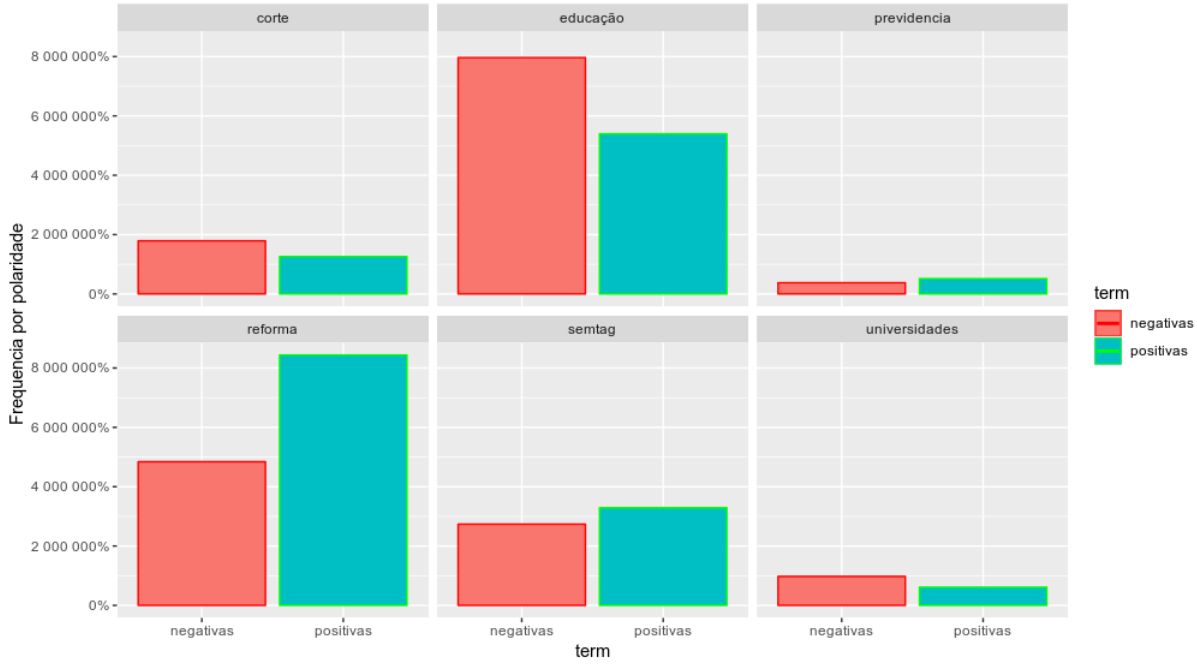

Figura 15: Gráfico de barras da frequência relativa de cada *hashtag*.

Com isso é possível determinar o sentimento da maioria das pessoas que postaram *tweets* relacionados a cada *hashtag* de forma isolada, bem como daqueles que estão relacionados às palavras, mas sem tag. Esses são apresentados na Tabela 5.

Tabela 5: Tabela do sentimento de cada *hashtag*.

Hashtag	Sentimento
corte	negativo
educação	negativo
previdência	positivo
reforma	positivo
sem tag	negativo
universidade	negativo

Já com relação ao valor total dos *tweets*, foi feita uma nuvem de palavras com os termos mais frequentes, indicando qual a polaridade de cada um, como pode-se observar na Figura 16. Tem-se a cor rosa e a parte superior para as palavras de polaridade negativa e palavras em azul na parte inferior para as positivas.

Já a Figura 17 mostra o quanto cada uma dessas palavras presentes contribuíram para a determinação do sentimento geral das pessoas que publicaram suas opiniões referentes ao assunto em questão. Neste gráfico é possível constatar que existe uma grande quantidade de palavras contribuindo para o sentimento negativo de forma isolada, ou seja, várias palavras com polaridade negativa, mas citadas poucas vezes em relação à quantidade total de *tweets*.

Já para o sentimento positivo o cenário é inverso, existem menos palavras que contribuem, no entanto elas aparecem com muito mais frequência, ou seja, elas aparecem repetidas vezes. Isso pode explicar o sentimento geral das pessoas, definido positivo, como pode ser visualizado na Figura 18. Percebe-se que esse sentimento é definido por aproximadamente o dobro da quantidade de palavras em relação ao sentimento negativo.

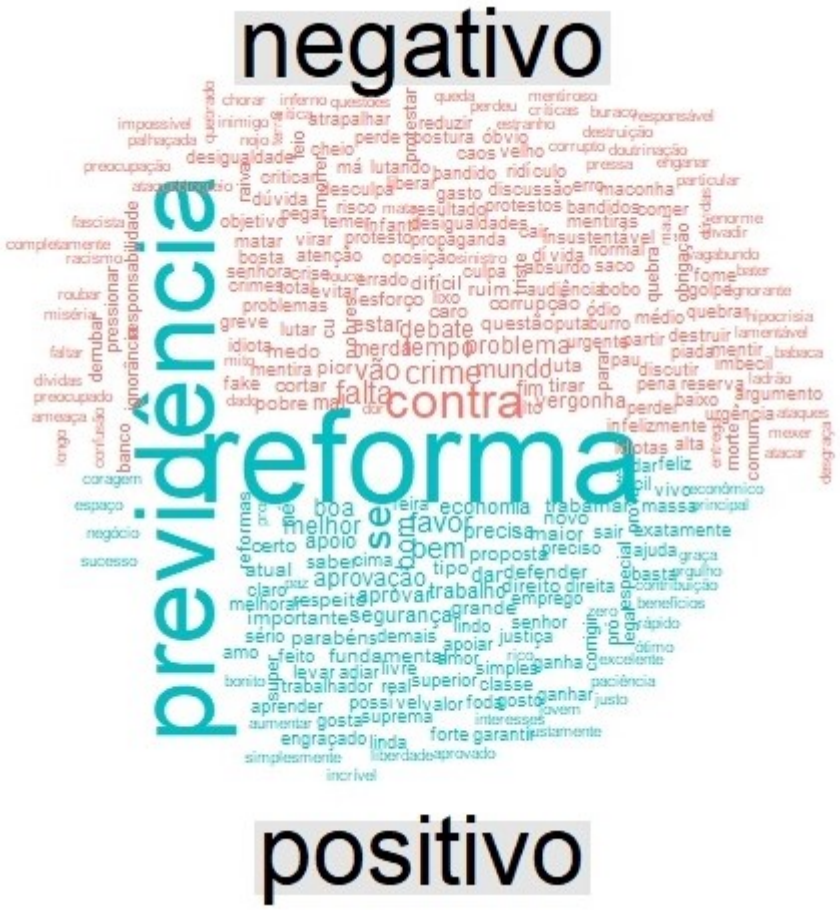

Figura 16: Nuvem de palavras com indicação do sentimento.

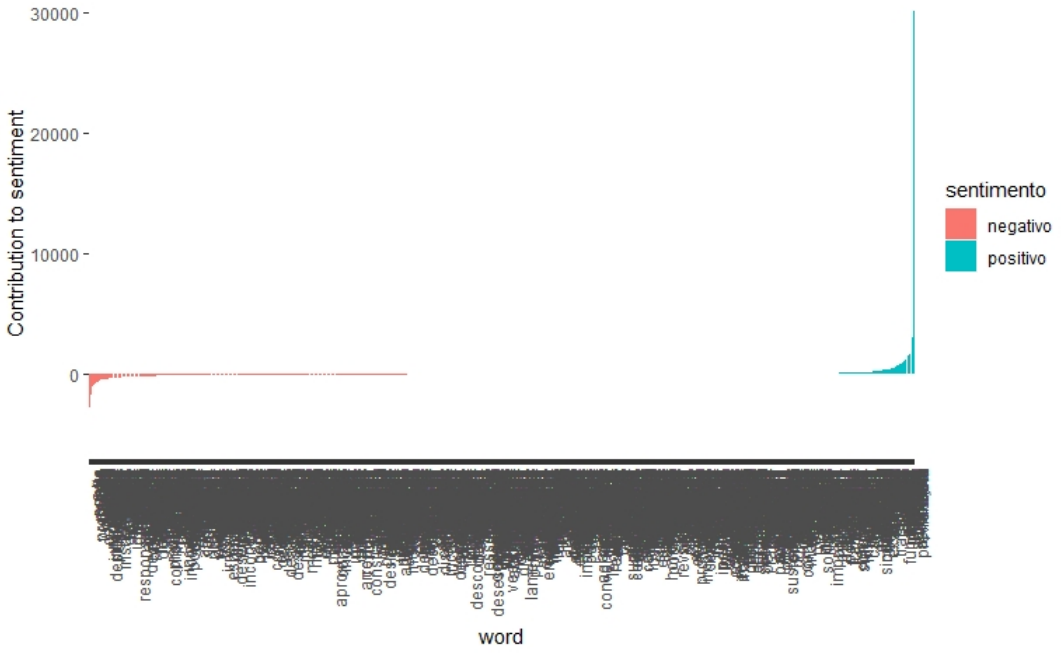

Figura 17: Gráfico da relevância de palavras para cada sentimento.

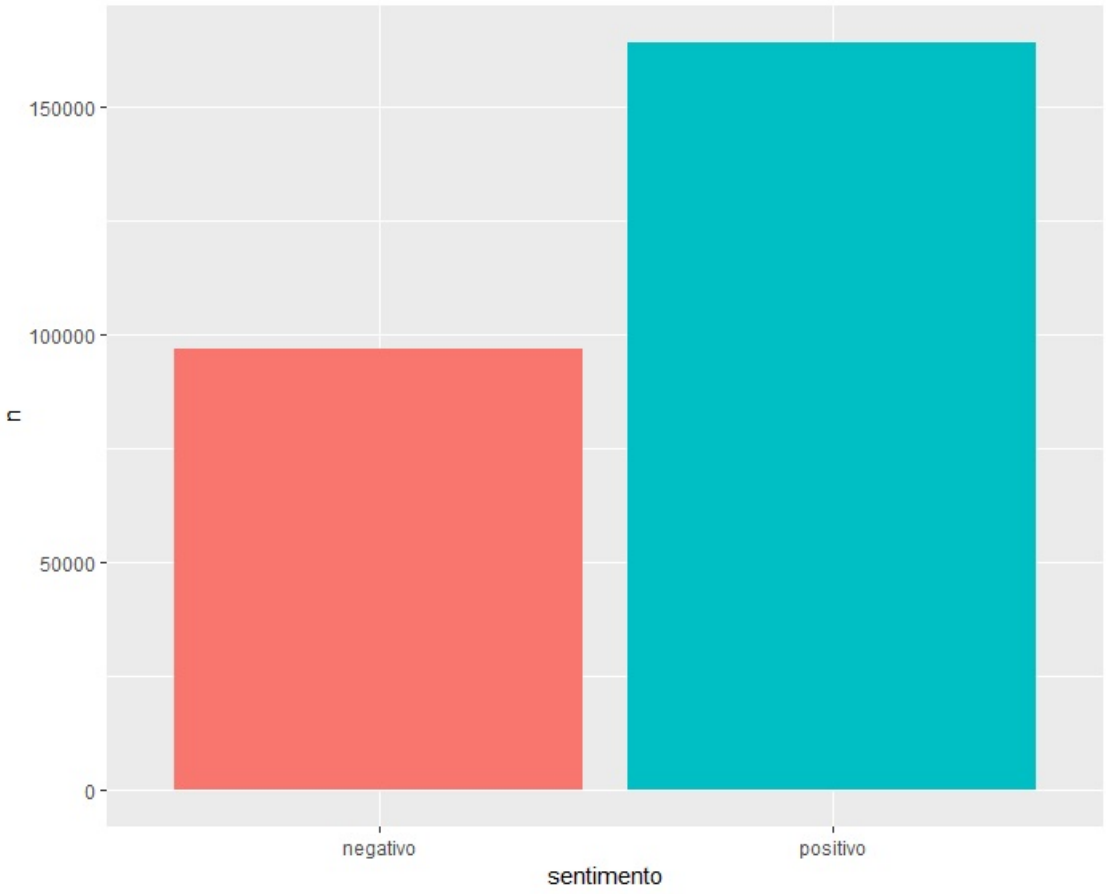

Figura 18: Gráfico de barras dos sentimentos.

O resultado obtido já era esperado, uma vez que na análise dos gráficos e tabelas obtidos anteriormente, a maioria das palavras e expressões tendiam de fato a serem a favor do atual governo brasileiro.

Com isso, teoricamente, tem-se que a maioria das pessoas possuem um sentimento positivo em relação à atual situação política do Brasil. No entanto, não se pode deixar de citar o fato de que os *tweets* a favor do atual presidente, Jair Bolsonaro, estavam muito em alta uma vez que dois dias depois aconteceriam as manifestações em todo o país apoiando o presidente e suas decisões para o futuro, como a reforma da previdência e o corte de verba na educação. Além disso, o presidente utiliza bastante o Twitter, fazendo com que seus apoiadores utilizem também.

Os comentários negativos se apresentaram de forma isolada, em uma intensidade muito menor, mas talvez se a coleta fosse feita mais próximo às manifestações a favor da educação e contra a reforma da previdência, que viriam a acontecer dias depois, esse resultado pudesse se alterar.

É necessário deixar claro que todos os fatos e acontecimentos afetam diretamente na quantidade de *tweets* falando a respeito do assunto, podendo em consequência afetar também as análises.

Um outro ponto em questão é de onde vêm esses *tweets*, isto é, qual a localização deles. Alguns usuários deixam acessível a localização e com isso, pôde-se perceber, através da Figura 19, que a maioria está concentrada em regiões onde o presidente foi eleito, ou seja, onde a maioria das pessoas são de fato a favor do atual governo e suas possíveis decisões, como na região Sudeste, por exemplo. Essa seria mais uma confirmação de que o resultado obtido tem grandes chances de ser verdadeiro.

Além disso existem *tweets* vindo de outros países, inclusive bem distantes do Brasil, o que talvez não faça muito sentido e possa ser um indício de possíveis robôs impulsionando determinados assuntos, o que pode vir a ser estudado posteriormente em trabalhos futuros.

Assim sendo, todas essas evidências são suposições para possíveis interpretações do resultado obtido.

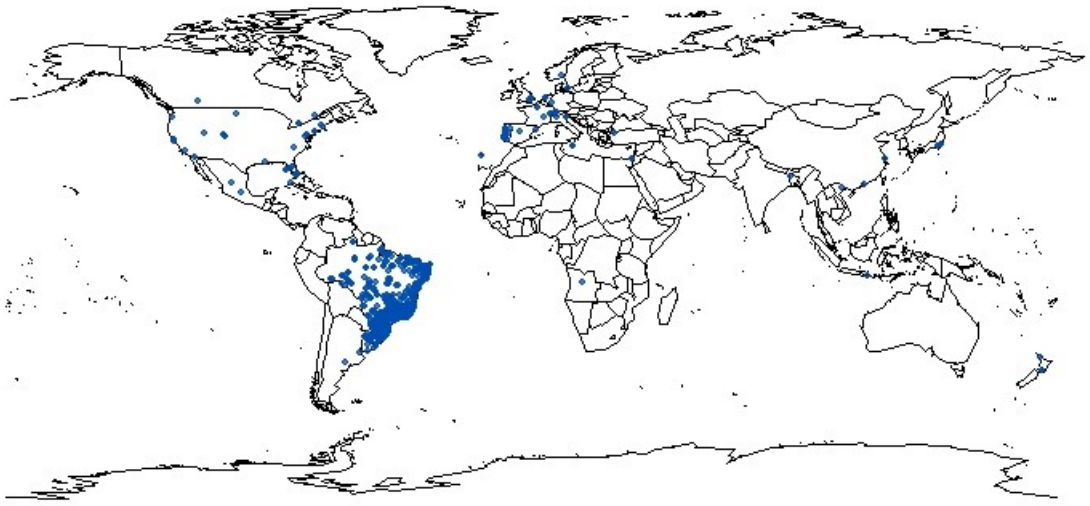

Figura 19: Mapa da localização dos usuários autores dos *tweets*.

Para finalizar este trabalho, também foi feito um gráfico interativo e em tempo real, estando parte dele apresentado na Figura 20, onde está exposta a localização de cada seguidor do presidente para que seja evidenciada a quantidade em cada local. Isso foi feito

através do ID de cada usuário e sua respectiva localização, quando disponibilizada pelo mesmo, no dia 06 de junho de 2019, às 18h.

Portanto, nota-se por exemplo, que existiam 2526 seguidores do presidente no Brasil, sendo 168 em Minas Gerais, 63 em Belo Horizonte, nesse dia e nesse horário. Essas informações variam de acordo com o momento, uma vez que ele pode ganhar e perder seguidores a todo tempo.

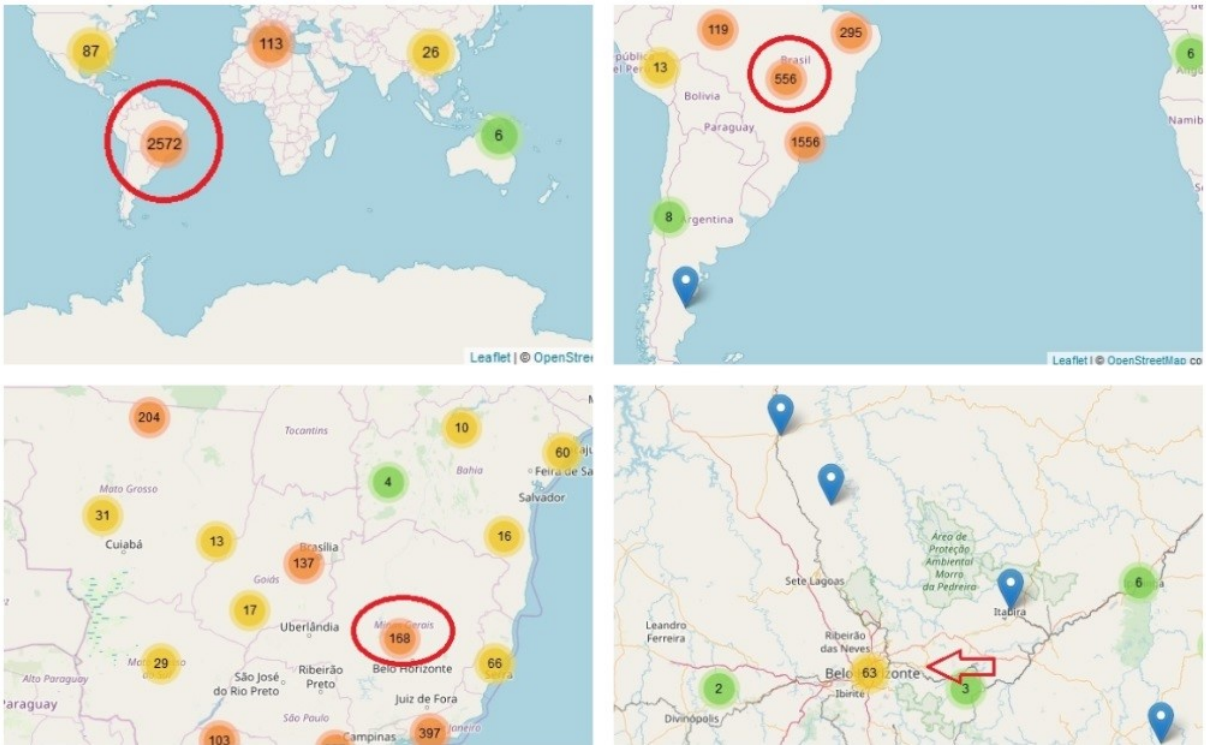

Figura 20: Mapa da localização dos seguidores do atual presidente do Brasil.

6 Considerações finais

Este trabalho demonstrou o estudo dos principais conceitos e técnicas relacionadas ao processo de descoberta do conhecimento através de Mineração de Dados e apresentou, em forma de aplicação, como o processo de Mineração de Dados foi usado para coletar, pré-processar o texto extraído do Twitter (*tweets*) e analisar os sentimentos nos quais os usuários desse *microblog* se baseavam ao escrever os *tweets* que falavam sobre a atual situação política do Brasil, o que permitiu conhecer a opinião dos usuários do Twitter sobre este assunto. As postagens dos usuários, popularmente conhecido como *tweets*, foram categorizadas neste trabalho em um sentimento positivo ou negativo.

Pode-se observar através desse trabalho, que os usuários dessa rede social apresentaram, em sua maioria, sentimento positivo quanto à reforma da previdência e sentimento negativo quanto ao corte de verba na educação, quando analisados de forma isolada.

Analisando de forma geral, o sentimento positivo predomina entre a maioria das pessoas com relação à atual situação política do Brasil. Sendo assim, teoricamente a maior parte da população brasileira está gostando do atual governo do país e de suas propostas de mudanças para o futuro. Entretanto, é importante ressaltar que dois dias depois que foi realizada a coleta dos dados, aconteceria uma manifestação a favor do atual presidente do Brasil e de sua política, o que pode ter influenciado na análise.

Quando feita a análise a respeito da localização de origem dos *tweets*, surgiu a possibilidade de existência de robôs que poderiam estar impulsionando publicações, mas esse estudo não foi feito à fundo, sendo um tema para possíveis trabalhos futuros.

Referências

ARMANO, G.; FANNI, F.; GIULIANI, A. Stopwords identification by means of characteristic and discriminant analysis. In: SCITEPRESS-SCIENCE AND TECHNOLOGY PUBLICATIONS, LDA. *Proceedings of the International Conference on Agents and Artificial Intelligence-Volume 2*. [S.l.], 2015. p. 353–360.

BERRY, M. J.; LINOFF, G. Data mining techniques: for marketing, sales, and customer support 1997. *John Willey & Sons*, 1997.

BLOMBERG, L. C. et al. Gestão de métricas e indicadores de doenças em saúde bucal suportado por um ambiente de descoberta de conhecimento em banco de dados. Pontifícia Universidade Católica do Rio Grande do Sul, 2010.

BRAGA, I. Avaliação da influência da remoção de stopwords na abordagem estatística de extração automática de termos. In: *7th Brazilian Symposium in Information and Human Language Technology (STIL 2009), So Carlos, SP, Brazil*. [S.l.: s.n.], 2009. p. 18.

BRAMER, M. *Principles of data mining*. [S.l.]: Springer, 2007.

CARDOSO, O. N. P.; MACHADO, R. T. M. Gestão do conhecimento usando data mining: estudo de caso na universidade federal de lavras. *Revista de administração pública*, v. 42, n. 3, p. 495–528, 2008.

CARVALHO, D. R.; ESCOBAR, L. F. A.; TSUNODA, D. Pontos de atenção para o uso da mineração de dados da saúde. *Informação & Informação*, v. 19, n. 1, p. 249–272, 2014.

CHEN, H. *Knowledge management systems: a text mining perspective*. [S.l.]: Knowledge Computing Corporation, 2001.

CORRÊA, A. C. G. et al. Recuperação de documentos baseados em informação semântica no ambiente ammo. Universidade Federal de São Carlos, 2003.

DENECKE, K. Using sentiwordnet for multilingual sentiment analysis. In: IEEE. *2008 IEEE 24th International Conference on Data Engineering Workshop*. [S.l.], 2008. p. 507–512.

FRAWLEY, W.; PIATETSKY-SHAPIRO, G.; MATHEUS, C. An overview: Knowledge discovery in databases. *Knowledge Discovery in Databases*, AAAI/MIT Press, p. 1–30, 1991.

FREITAS, C. Sobre a construção de um léxico da afetividade para o processamento computacional do português. *Revista Brasileira de Linguística Aplicada*, SciELO Brasil, v. 13, n. 4, 2013.

GOLDSCHMIDT, R.; BEZERRA, E.; PASSOS, E. Data mining: Conceitos, técnicas, algoritmos, orientações e aplicações. *Rio de Janeiro-RJ: Elsevier*, 2015.

HAN, J.; KAMBER, M. *Data Mining: Concepts and Techniques. A. Stephan.* [S.l.]: San Francisco,, Morgan Kaufmann Publishers is an imprint of Elsevier, 2006.

HEKIMA. 2014. (Como classificar sentimentos nas redes sociais?).

LAROSE, D. T. An introduction to data mining. *Discovering Knowledge in Data, John Wiley & Sons Publication, Hoboken, New Jersey, USA*, p. 1–25, 2005.

LIU, B. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, Morgan & Claypool Publishers, v. 5, n. 1, p. 1–167, 2012.

MARISCAL, G.; MARBAN, O.; FERNANDEZ, C. A survey of data mining and knowledge discovery process models and methodologies. *The Knowledge Engineering Review*, Cambridge University Press, v. 25, n. 2, p. 137–166, 2010.

MEYFROIDT, G. et al. Machine learning techniques to examine large patient databases. *Best Practice & Research Clinical Anaesthesiology*, Elsevier, v. 23, n. 1, p. 127–143, 2009.

NL, R. 2018. (Brazilian Stopwords).

OLSON, D. L.; DELEN, D. *Advanced data mining techniques.* [S.l.]: Springer Science & Business Media, 2008.

PANG, B.; LEE, L. et al. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, Now Publishers, Inc., v. 2, n. 1–2, p. 1–135, 2008.

PASSOS, E.; ARANHA, C. A tecnologia de mineração de textos. *RESI-Revista Eletrônica de Sistemas de Informação*, v. 2, 2006.

SILVA, N. F. F. d. *Análise de sentimentos em textos curtos provenientes de redes sociais.* Tese (Doutorado) — Universidade de São Paulo, 2016.

TAN, A.-H. et al. Text mining: The state of the art and the challenges. In: SN. *Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases.* [S.l.], 1999. v. 8, p. 65–70.

TATMAN, R. 2017. (Sentiment Lexicons for 81 Languages).

TEAM, R. C. R foundation for statistical computing; vienna, austria: 2014. *R: A language and environment for statistical computing*, p. 2013, 2018.

TRENDS, G. 2019. (Veja o que o mundo está pesquisando).

TSYTSARAU, M.; PALPANAS, T. Survey on mining subjective data on the web. *Data Mining and Knowledge Discovery*, Kluwer Academic Publishers, v. 24, n. 3, p. 478–514, 2012.

WANG, J.; HU, X.; ZHU, D. Data mining in public administration. In: *Handbook of Research on Public Information Technology.* [S.l.]: IGI Global, 2008. p. 556–567.

WRITTEN, I.; FRANK, E. *Data Mining: Practical Machine Learning Tools and Techniques, Second Edi.* [S.l.]: Elsevier Science, 2005.

ANEXO A – Código na Linguagem R

#Pacotes necessários

options(max.print=5.5E5)

carregar <- function(pkg) {
new.pkg <- pkg{(pkg %in% installed.packages()[, ”Package”])]
if (length(new.pkg))
install.packages(new.pkg, dependencies = TRUE)
sapply(pkg, require, character.only = TRUE) }

packages <- c(”twitteR”, ”tidyverse”, ”data.table”, ”tidytext”, ”glue”, ”stringr”, ”stringi”, ”rvest”, ”readr”, ”ptstem”, ”wordcloud2”, ”wordcloud”, ”tm”, ”RColorBrewer”, ”igraph”, ”quanteda”, ”topicmodels”, ”sentR”, ”SentimentAnalysis”, ”RJSONIO”, ”leaflet”, ”scales”, ”ggplot2”, ”reshape2”, ”rtweet”, ”ggmap”)

carregar(packages)

#Carregando as stopwords

stopwords = read.table("stopwords.txt",colClasses = ’character’)

str(stopwords)

stw=unlist(str_split(stopwords$V1,’ \\n’))

glimpse(stw)

#Recuperando stopwords em português do TM

sw=tm::stopwords(kind = "portuguese")

glimpse(sw)

#Fazendo um merge nos dados

sw_merged <- union(sw,stw)

#Vamos dar uma verificada se deu tudo certo, se não existem elementos repetidos na tabela. Para isso uso o objeto tibble do dplyr

```
dplyr::tibble(word = sw_merged) %>%
group_by(word) %>%
filter(n()>1)
```

#Vendo as 6 primeiras linhas da lista de stopwords

```
head(sw_merged)
```

#Carregando os termos de polaridade de Análise de Sentimentos em uma tabela geral

```
an <- read.csv("adjetivos_negativos.txt", header = F, sep = "\t", strip.white = F, stringsAsFactors = F, encoding="UTF-8")
exn <- read.csv("expressoes_negativas.txt", header = F, sep = "\t", strip.white = F, stringsAsFactors = F, encoding="UTF-8")
vn <- read.csv("verbos_negativos.txt", header = F, sep = "\t", strip.white = F, stringsAsFactors = F, encoding="UTF-8")
subn <- read.csv("substantivos_negativos.txt", header = F, sep = "\t", strip.white = F, stringsAsFactors = F, encoding="UTF-8")
ap <- read.csv("adjetivos_positivos.txt", header = F, sep = "\t", strip.white = F, stringsAsFactors = F, encoding="UTF-8")
exp <- read.csv("expressoes_positivas.txt", header = F, sep = "\t", strip.white = F, stringsAsFactors = F, encoding="UTF-8")
vp <- read.csv("verbos_positivos.txt", header = F, sep = "\t", strip.white = F, stringsAsFactors = F, encoding="UTF-8")
sp <- read.csv("substantivos_positivos.txt", header = F, sep = "\t", strip.white = F, stringsAsFactors = F, encoding="UTF-8")
```

#Carregando o léxico de sentimentos disponível no Kaggle

#from kaggle sentiment words https://www.kaggle.com/rtatman/sentiment-lexicons-for-81-languages/data

```
poskaggle <- read.csv("positive_words_pt.txt", header = F, sep = "\t", strip.white = F, stringsAsFactors = F, encoding="UTF-8")
negkaggle <- read.csv("negative_words_pt.txt", header = F, sep = "\t", strip.white = F, stringsAsFactors = F, encoding="UTF-8")
```

#Verificando se está tudo certo

```
head(poskaggle)
```

```
head(negkaggle)
```

#Criando um dataframe com os termos, começando pelos adjetivos negativos

```
dfPolaridades <- an %>%
mutate(word = V1, polaridade = -1, tipo='adjetivo', sentimento='negativo') %>%
select(word,polaridade,tipo,sentimento) %>%
arrange(word)
head(dfPolaridades,2)
```

#Aqui faço um count para poder adicionar os dados corretamente

```
icount <- length(vn$V1)
dfPolaridades <- bind_rows(dfPolaridades, list(word=vn$V1, polaridade=rep(-1, i-
count), tipo = rep('verbo',icount),sentimento = rep('negativo',icount)))
icount <- length(subn$V1)
dfPolaridades <- bind_rows(dfPolaridades,list(word = subn$V1, polaridade=rep(-1,
icount), tipo=rep('substantivo',icount),sentimento=rep('negativo',icount)))
icount <- length(negkaggle$V1)
dfPolaridades <- bind_rows(dfPolaridades, list(word = negkaggle$V1, polaridade =
rep(-1,icount), tipo=rep('noclass',icount),sentimento=rep('negativo',icount)))
icount <- length(ap$V1)
dfPolaridades <- bind_rows(dfPolaridades,list(word = ap$V1, polaridade=rep(1, i-
count), tipo=rep('adjetivo',icount),sentimento=rep('positivo',icount)))
icount <- length(exp$V1)
dfPolaridades <- bind_rows(dfPolaridades,list(word = exp$V1,polaridade=rep(1, i-
count), tipo=rep('expressao',icount),sentimento=rep('positivo',icount)))
icount <- length(vp$V1)
dfPolaridades <- bind_rows(dfPolaridades, list(word = vp$V1, polaridade=rep(1, i-
count), tipo=rep('verbo',icount),sentimento=rep('positivo',icount)))
icount <- length(sp$V1)
dfPolaridades <- bind_rows(dfPolaridades,list(word = sp$V1, polaridade=rep(1, i-
count), tipo=rep('substantivo',icount),sentimento=rep('positivo',icount)))
icount <- length(poskaggle$V1)
dfPolaridades <- bind_rows(dfPolaridades,list(word = poskaggle$V1, polaridade =
rep(1,icount), tipo=rep('noclass',icount),sentimento=rep('positivo',icount)))
```

#Visualizando como está nosso dataframe

dfPolaridades %>% group_by(word) %>% filter(n() == 1) %>%
summarize(n=n())
dfPolaridades %>% count()

#Removendo termos de sentimentos repetidos

dfPolaridadesUnique <- dfPolaridades[!duplicated(dfPolaridades$word),]
dfPolaridadesUnique %>% count()
tibble(dfPolaridadesUnique)

#Conectando e autorizando (aqui são necessárias as informações fornecidas após criado um app gratuito no site developer do Twitter)

#Obs.: é necessário ter uma conta no Twitter para obter os códigos a seguir

api_key <- "**API key** (codigo numerico)"
api_secret <- "**API secret key** (codigo numerico)"
access_token <- "**Access token** (codigo numerico)"
access_token_secret <- "**Access token secret** (codigo numerico)"
setup_twitter_oauth(api_key,api_secret)

#Verificando trends

trendsBR <- getTrends(woeid = 23424768)

#10 primeiros apenas

trendsBR$name[1:10]

#Definindo as palavras e hashtags que serão utilizadas na pesquisa pelos 100 mil tweets

tweets <- search_tweets(q = "reforma OR previdencia OR universidades OR corte OR educacao OR #reformadaprevidencia OR #universidades OR #corte OR #educacao", retryonratelimit = TRUE, lang = "pt", n = 100000, include_rts = FALSE, type = 'mixed')
save(tweets, file="tweets.RData") #comando utilizado para salvar os tweets obtidos
#load("tweets.RData") comando utilizado caso se queira carregar os tweets já salvos
dim(tweets)
head(tweets)
tweetxt <- tweets$text
tibble(tweetxt)

```
#Processo de limpeza

removeURL <- function(x) gsub("http[ ^ [:space:]]*"," ", x)
tweetxtUtf <- readr::parse_character(tweetxt, locale = readr::locale('pt'))
#sapply(tweetxt, function(x) iconv(x,"UTF-8"))
tweetxtUtf <- sapply(tweetxtUtf, function(x) stri_trans_tolower(x,'pt'))
tweetxtUtf <- gsub("(RT|via)((?:\\b\\W*@\\w+)+)", " ", tweetxtUtf);
tweetxtUtf <- str_replace(tweetxtUtf,"RT @[a-z,A-Z]*: ","")
tweetxtUtf <- gsub("@ \\w+", " ", tweetxtUtf)
tweetxtUtf <- removeURL(tweetxtUtf)
tweetxtUtf <- str_replace_all(tweetxtUtf,"@[a-z,A-Z]*"," ")
tweetxtUtf <- gsub("[ ^ [:alnum:][:blank:]!?]", " ", tweetxtUtf)
tweetxtUtf <- gsub("[[:digit:]]", " ", tweetxtUtf)
tibble(tweetxtUtf)

#Não esquecer de remover os tweets repetidos (vamos ver)

length(tweetxtUtf)
tibble(tweetxtUtf) %>% unique() %>% count()
tweetxtUtfUnique <- tweetxtUtf %>% unique()
length(tweetxtUtfUnique)
tweetxtUtfUniqueSw <- tm::removeWords(tweetxtUtfUnique,c(sw_merged,'rt'))
tibble(tweetxtUtfUniqueSw)

#Nuvem de palavras

ttokens <- tibble(word= tweetxtUtfUniqueSw) %>% unnest_tokens(word,word)
ttokens %>% count(word, sort = T)
ttokens_filter <- ttokens %>% filter(nchar(word) > 3)
ttokens_filter %>% count(word, sort=T)
ttokens_freq <- ttokens_filter %>% count(word, sort = T) %>% select(word, freq=n)
ttokens_freq
wordcloud2(ttokens_freq , minSize = 2, size = 2)
pal2 <- brewer.pal(8,"Dark2")
wordcloud(words = ttokens_freq$word, freq = ttokens_freq$freq , min.freq = 8, ran-
dom.color = T, max.word = 200, random.order = T, colors = pal2)

#Criando o corpus

tweetencoded <- sapply(tweetxtUtfUnique,enc2native)
```

```
df <- data.frame(text=tweetencoded)
head(df)
```

#Criando um id para o documento

```
df$doc_id <- row.names(df)
head(df)
tm_corpus <- Corpus(DataframeSource(df))
inspect(tm_corpus[1:4])
```

#Criando o dtm usando palavras que tem mais de 3 letras e uma frequencia>19

```
dtm <- DocumentTermMatrix(tm_corpus, control=list(wordLengths=c(4, 20),language=locale('pt'), stopwords=stopwords('portuguese'), bounds = list(global=c(30,500))))
dtm
```

#Termos mais frequentes

```
findFreqTerms(dtm)
twdf <- tibble(tweet = tweetxtUtfUnique)
twdf$whois <- NA
twdf$whois[twdf$tweet %like% 'reforma'] <- 'reforma'
twdf$whois[twdf$tweet %like% 'previdencia'] <- 'previdencia'
twdf$whois[twdf$tweet %like% 'universidades'] <- 'universidades'
twdf$whois[twdf$tweet %like% 'corte'] <- 'corte'
twdf$whois[twdf$tweet %like% 'educacao'] <- 'educacao'
twdf$whois[is.na(twdf$whois) ] <- 'semtag'
freq <- twdf %>% count(whois, sort = T) %>% select( whois,freq = n)
freq
pie(table(twdf$whois))  #Gráfico de setores
barplot(table(twdf$whois))  #Gráfico de barras
```

#Criando o dtm

```
dfq <- data.frame(id=row.names(twdf), text=twdf$tweet, whois=factor(twdf$whois))
myCorpus <- corpus(twdf, text_field = 'tweet', metacorpus = list(source = "tweets do governo"))
myCorpus
```

#Criando um gráfico das relações entre as tags

twraw <- readr::parse_character(tweetxt, locale = readr::locale('pt'))

mytoken <- tokens(twraw, remove_numbers=T,remove_symbols=T, remove_twitter =T, remove_url=T)

head(mytoken)

mytoken <- tokens_remove(mytoken, stopwords('portuguese'))

head(textstat_collocations(mytoken,size = 5, min_count = 5))

myrawCorpus <- corpus(twraw)

tweetdfm <- dfm(myrawCorpus, remove_punct = TRUE)

tagdfm <- dfm_select(tweetdfm, ('#*'))

toptag <- names(topfeatures(tagdfm, 50))

head(toptag)

tagfcm <- fcm(tagdfm)

head(tagfcm)

toptagfcm <- fcm_select(tagfcm, toptag)

textplot_network(toptagfcm, min_freq = 0.1, edge_alpha = 0.8, edge_size = 5)

#Termos mais frequentes

myDfm <- dfm(myCorpus, stem = F)

myDfm

topfeatures(myDfm,20)

stopwors2 <- c('the','r','Ã©','c','?','!','of','rt','pra') myDfm<-dfm(myCorpus,groups ='whois',remove=c(quanteda::stopwords("portuguese"), stopwors2,tm::stopwords('portuguese')), stem = F,remove_punct = TRUE)

#Note que com a opção groups ele agrupou pela nossa classificação

myDfm

#Para acessar os termos mais usados

topfeatures(myDfm, 20)

allfeats <- textstat_frequency(myDfm)

allfeats$feature <- with(allfeats, reorder(feature, -frequency))

ggplot(head(allfeats,20), aes(x=feature,y=frequency,fill=frequency)) + geom_bar(stat="identity") + xlab("Termos") + ylab("Frequência") + coord_flip() + theme(axis.text =element_text(size=7)) **#Gráfico dos termos mais frequentes**

col <- textstat_collocations(myCorpus , size = 2:4, min_count = 2)

head(col)

ggplot(col[order(col$count, decreasing = T),][1:25,], aes(x=reorder(collocation,count), y=factor(count), fill=factor(count))) + geom_bar(stat="identity") + xlab("Expressões") + ylab("Frequência") + coord_flip() + theme(axis.text=element_text(size=7)) **#Gráfico das expressões mais frequentes**

#Criando um DTM com nossas polaridades como dicionário

positivas <- as_tibble(dfPolaridadesUnique) %>% filter(sentimento == 'positivo') %>% select(word)

negativas <- as_tibble(dfPolaridadesUnique) %>% filter(sentimento == 'negativo') %>% select(word)

dic <- dictionary(list(positivas =as.character(positivas$word), negativas = as.character(negativas$word)))

bySentimento <- dfm(myCorpus, dictionary = dic)

scorebygroup <- tidy(bySentimento %>% dfm_group(groups='whois'))

scorebygroup

scorebygroup %>% ggplot(aes(document, count)) + geom_point() + geom_smooth() + facet_wrap(~ term) + scale_y_continuous(labels = percent_format()) + ylab("Frequência por polaridade") + aes(color = term) + scale_color_manual(values = c("red", "green"))

scorebygroup %>% ggplot(aes(term, count)) + geom_bar(stat = 'identity') + geom_smooth() + facet_wrap(~ document) + scale_y_continuous(labels = percent_format()) + ylab("Frequência por polaridade") + aes(fill= term,color = term) + scale_color_manual(values = c("red", "green"))

#Análise de Sentimentos

#Análise de Sentimentos com Tidy

#Salvando o dataframe para não ter discrepâncias nos comentários

#twdf %>% write_csv(path='twittersentimentaldata.csv')

#twdf <- read_csv('twittersentimentaldata.csv')

twdf$id <- rownames(twdf)

tw <- twdf %>% mutate(document = id, word = tweet) %>% select(document, word, whois)

#Note que a coluna document carrega a identificação de cada texto

str(tw)

tdm <- tw %>% unnest_tokens(word,word)

#Removendo as stopwords

tdm <- tdm %>% anti_join(data.frame(word=stopwords('portuguese'))
head(tdm)
#tdm <- tdm %>% group_by(document) %>% mutate(word_per_doc = n())
#tdm <- tdm %>% group_by(whois) %>% mutate(word_per_whois = n())

#Plotando as primeiras impressões

library(tidyr)
sentJoin <- tdm %>% inner_join(as_tibble(dfPolaridadesUnique), by='word')
sentJoin %>% count(sentimento) %>% ggplot(aes(sentimento,n , fill = sentimento)) + geom_bar(stat = "identity", show.legend = FALSE)

#Quais são nossos top sentimentos?

word_counts <- sentJoin %>% count(word, sentimento, sort = TRUE) %>% ungroup()
word_counts %>% filter(n > 5) %>% mutate(n = ifelse(sentimento == "negativo", -n, n)) %>% mutate(word = reorder(word, n)) %>% ggplot(aes(word, n, fill = sentimento)) + geom_bar(stat = "identity") + theme(axis.text.x = element_text(angle = 90, hjust = 1)) + ylab("Contribution to sentiment")

#E finalmente nossa nuvem de palavras de sentimentos

sentJoin %>% count(word, sentimento, sort = TRUE) %>% acast(word ~ sentimento, value.var = "n", fill = 0) %>% comparison.cloud(colors = c("#F8766D", "#00BFC4"), max.words = 300)

#Sentimentos mais negativos

bottom8tw <- head(sentJoin %>% count(document, sentimento) %>% spread(sentimento, n, fill = 0) %>% mutate(score = positivo - negativo) %>% arrange(score),8)['document']
twdf %>% filter(id %in% as.vector(bottom8tw$document))

#Sentimentos mais positivos

top8 <- head(sentJoin %>% count(document, sentimento) %>% spread(sentimento, n, fill = 0) %>% mutate(score = positivo - negativo) %>% arrange(desc(score)),8)[' document']
twdf %>% filter(id %in% as.vector(top8$document))

```
#Salvando

write_csv(tibble(word = sw_merged),path = 'stopwords.csv')
write_csv(as.tibble(dfPolaridadesUnique),path = 'polaridades_pt.csv')
write_csv(tibble(tweet = tweetxt),path = 'tweetxt.csv')
write_csv(tibble(tweet = tweetxtUtfUniqueSw),path='tweets_limpo.csv')
write_csv(ttokens_freq, path='tokens.csv')

#Plotando o mapa do mundo

geo_tweet <- lat_lng(tweets)
par(mar = c(0,0,0,0))
maps::map("world", lwd = 0.25)
with(geo_tweet, points(lng, lat, pch = 20, cex = .75, col = rgb(0, .3, .7, .75)))

#SEGUNDA PARTE: SEGUIDORES DA CONTA DO PRESIDENTE

#Inserindo o usuário que você deseja pesquisar

usuario <- "jairbolsonaro"

#Baixando os dados desse usuário

dados_usuario <- lookup_users(usuario)
dados_usuario$location

#Buscando os ids dos seguidores do usuário
#followers_usuario <-get_followers(usuario, n = dados_usuario$followers_count, retryonratelimit = TRUE)
followers_usuario <-get_followers(usuario, n = 10000, retryonratelimit = TRUE)

#Observando a quantidade de dados de seguidores que foi buscado

length(followers_usuario$user_id)
followers_usuario$user_id

#Buscando dados dos seguidores

dados_usuario_followers <- lookup_users(followers_usuario$user_id)
str(dados_usuario_followers)

#Verificando as primeiras 10 localizações buscadas no passo acima

head(dados_usuario_followers$location, 10)

#Reescrevendo o dataframe apenas com usuários com localização
```

dados_usuario_followers <-subset(dados_usuario_followers, location!=)
head(dados_usuario_followers)

#Criando o dataframe de resultados

geocode_results <- dados_usuario_followers[1:length(dados_usuario_followers$user_id),c(3,4)]

#Inserindo a barra de progresso para vizualização do processo

pb <- txtProgressBar(min = 0, max = nrow(geocode_results), style = 3)

#Buscando pelos geocodes (latitude/longitude)

for(i in 1:nrow(geocode_results)) {
#print("Buscando...")
result<-geocode(dados_usuario_followers$location[i],output="latlona", source="dsk")
geocode_results$lat[i] <- as.numeric(result[2])
geocode_results$lon[i] <- as.numeric(result[1])
Sys.sleep(0.1)
setTxtProgressBar(pb, i) }

#Fechando a barra de progresso

close(pb)

#Criando um rótulo de identificação do usuário para ser usado no mapa

bio_html <- as.data.frame(paste(«b> <img src='", dados_usuario_followers$profile_image_url,"'height='64" width='64' > <a href ='http://www.twitter.com/",dados_usuario_followers$screen_name,"' target='_blank' >", dados_usuario_followers$name,«/a> ", «p> Localização:
 ",dados_usuario_followers$location,«/center> </p>", sep = " ")

#Renomeando esse dataframe de rótulo

names(bio_html) <- "bio"

#Criando um dataframe final com todos os dados para serem plotados no mapa

geocode_data <- cbind.data.frame(geocode_results,bio_html)

#Plotando as geolocalizões dos usuários em um mapa interativo

geocode_data %>%

leaflet() %>% #**carrega o leaflet**

addTiles() %>% #**adiciona as camadas de mapas de acordo com o zoom**

addMarkers(lng = ~ lon, lat = ~ lat,popup= ~ bio, #**mapeia a base de dados de acordo com as respectivas lat e lon**

clusterOptions = markerClusterOptions()) #**cria clusters em zoom para melhor visualização**