



# **CENTRO UNIVERSITÁRIO LUTERANO DE PALMAS**

Recredenciado pela Portaria Ministerial nº 1.162, de 13/10/16, D.O.U. nº 198, de 14/10/2016  
AELBRA EDUCAÇÃO SUPERIOR - GRADUAÇÃO E PÓS-GRADUAÇÃO S.A.

[Leticia Bueno Martins](#)

EXTRAÇÃO DE ASPECTOS EM *TWEETS* PUBLICADOS POR MULHERES QUE  
ATUAM NA ÁREA DE TECNOLOGIA

Palmas – TO

2022

Leticia Bueno Martins

EXTRAÇÃO DE ASPECTOS EM *TWEETS* PUBLICADOS POR MULHERES QUE  
ATUAM NA ÁREA DE TECNOLOGIA

Projeto de Pesquisa elaborado e apresentado como requisito parcial para aprovação na disciplina de Laboratório de Criação do curso de bacharel em Engenharia de Software pelo Centro Universitário Luterano de Palmas (CEULP/ULBRA).

Orientador: Prof. Esp. Fábio Araujo Castro

Palmas – TO

2022

Leticia Bueno Martins

EXTRAÇÃO DE ASPECTOS EM *TWEETS* PUBLICADOS POR MULHERES QUE  
ATUAM NA ÁREA DE TECNOLOGIA

Projeto de Pesquisa elaborado e apresentado como requisito parcial para aprovação na disciplina de Laboratório de Criação do curso de bacharel em Engenharia de Software pelo Centro Universitário Luterano de Palmas (CEULP/ULBRA).

Orientador: Prof. Esp. Fábio Araujo Castro.

Aprovado em: \_\_\_\_/\_\_\_\_/\_\_\_\_

BANCA EXAMINADORA

---

Prof. Esp. Fábio Araujo Castro

Orientador

Centro Universitário Luterano de Palmas – CEULP

---

Prof. Dra. Parcilene Fernandes de Brito

Centro Universitário Luterano de Palmas – CEULP

---

Prof. Esp. Fernanda Pereira Gomes

Centro Universitário Luterano de Palmas – CEULP

Palmas – TO

2022

## RESUMO

MARTINS, Leticia Bueno. **Extração de Aspectos em Tweets Publicados Por Mulheres Que Atuam Na Área De Tecnologia**. 2022. 21 f. Projeto Tecnológico (Graduação) – Curso de Engenharia de Software, Centro Universitário Luterano de Palmas, Palmas/TO, 2022.

A baixa presença de mulheres na área de tecnologia da informação é um problema pautado no mercado de trabalho há muitos anos. Com o surgimento das redes sociais como maior ferramenta para comunicação e a quantidade de dados gerados, cria-se a oportunidade para pesquisadores coletar dados para realizar análises e investigar problemas. Este trabalho visa utilizar técnicas computacionais para realizar a extração de aspectos em *tweets* realizados por mulheres que atuam na área de tecnologia, criando uma possibilidade de identificar possíveis gargalos a respeito da presença da mulher na área de tecnologia da informação. A realização dessa extração é realizada em 3 etapas: coleta de dados, tratamento de *tweets* e extração de aspectos. Para a primeira etapa foi utilizada a API do Twitter para a coleta dos *tweets* e o banco de dados PostgreSQL para o armazenamento dos dados coletados. A segunda etapa foi realizada o processamento de linguagem natural e tratamento desses dados com o auxílio da biblioteca NLTK. Para a terceira etapa, a biblioteca spaCy foi utilizada para o auxílio da extração de aspectos dos *tweets*. Ao final do processo, é apresentado os aspectos extraídos

**PALAVRAS-CHAVE:** Extração de Aspectos, Processamento de Linguagem Natural, Twitter, Mulheres na Tecnologia.

## LISTA DE FIGURAS

<b>Figura 1</b> - Fluxo de aprendizagem supervisionado de máquina	16
<b>Figura 2</b> - Fluxo de aprendizagem não supervisionado baseado em léxicos	18
<b>Figura 3</b> - Fluxo de trabalho	20
<b>Figura 4</b> - Arquitetura do sistema de análise de sentimentos	23
<b>Figura 5</b> - Arquitetura do módulo de coleta de dados	24
<b>Figura 6</b> - Fluxograma de extração dos <i>tweets</i>	25
<b>Figura 7</b> - <i>Schema</i> do banco de dados	26
<b>Figura 8</b> - Fluxograma de tratamento de <i>tweets</i>	27
<b>Figura 9</b> - Exemplo de tokenização de palavras	29
<b>Figura 10</b> - Exemplo do algoritmo de correção ortográfica	30
<b>Figura 11</b> - Exemplo de sentença etiquetada	32
<b>Figura 12</b> - Exemplo de fluxo do algoritmo de extração de aspectos	33
<b>Figura 13</b> - Tabela com alguns aspectos selecionados	34
<b>Figura 14</b> - Gráfico com aspectos selecionados e sua frequência	35

## LISTA DE TABELAS

<b>Tabela 1</b> - Exemplo de normalização	28
<b>Tabela 2</b> - Exemplo de remoção de <i>stop-words</i>	29
<b>Tabela 3</b> - Exemplo de lematização	30
<b>Tabela 4</b> - Exemplo de correção ortográfica	31
<b>Tabela 5</b> - Exemplo de <i>tagset</i>	32
<b>Tabela 6</b> - Expressões regulares para normalização de <i>tweets</i>	42
<b>Tabela 7</b> - Expressões regulares para tratamento específico de emoticons	42
<b>Tabela 8</b> - Frequência dos aspectos coletados	43

## LISTA DE APÊNDICES

<b>APÊNDICE A</b> - EXPRESSÕES REGULARES DE TODAS AS REMOÇÕES E SUBSTITUIÇÕES REALIZADAS NA ETAPA DE TRATAMENTO DE <i>TWEETS</i>	42
<b>APÊNDICE B</b> - EXPRESSÕES REGULARES DE TODAS AS REMOÇÕES E SUBSTITUIÇÕES REALIZADAS NA ETAPA DE TRATAMENTO DE <i>TWEETS</i>	43

## **LISTA DE ABREVIATURAS E SIGLAS**

API - Application Programming Interface

MAXENT - Maxima Entropia

NB - Naive Bayes

NLTK – Natural Language Toolkit

PLN – Processamento de Linguagem Natural

PSF - Python Software Foundation

SQL – Structured Query Language

SVM - Support Vector Machine

TSA - Twitter Sentiment Analysis

## SUMÁRIO

<b>1 INTRODUÇÃO</b>	<b>9</b>
<b>2 REFERENCIAL TEÓRICO</b>	<b>12</b>
2.1 ANÁLISE DE SENTIMENTOS	12
2.2 ANÁLISE DE SENTIMENTOS EM PEQUENOS TEXTOS	13
<b>2.2.1 - Desafio em tratar tweets</b>	<b>14</b>
2.3 EXTRAÇÃO DE ASPECTOS	15
2.4 APRENDIZADO SUPERVISIONADO	17
2.5 APRENDIZADO SEMI-SUPERVISIONADO	18
2.6 APRENDIZADO NÃO SUPERVISIONADO BASEADO EM LÉXICO	18
<b>3 METODOLOGIA</b>	<b>21</b>
3.1 MATERIAIS	21
<b>4 RESULTADOS E DISCUSSÕES</b>	<b>24</b>
4.1 COLETA DE DADOS	26
4.2 - TRATAMENTO DE TWEETS	33
4.3 EXTRAÇÃO DE ASPECTOS	33
	36
<b>5 CONSIDERAÇÕES FINAIS</b>	<b>37</b>
<b>6 REFERÊNCIAS</b>	<b>38</b>
<b>7 APÊNDICES</b>	<b>43</b>

## 1 INTRODUÇÃO

Nos primeiros anos de suas carreiras na área de tecnologia, mulheres se autodenominam felizes e determinadas em suas profissões, mas, com o tempo, revelam enfrentar barreiras no local de trabalho (GARDNER, 2014). Ainda segundo Gardner (2014), 23% a 66% relatam experiências de assédio sexual ou viram isso acontecer a outras mulheres. Metade das entrevistadas indicaram que foram tratadas de uma forma que consideram hostil, humilhante ou condescendente, além disso, um terço também relatou que seus chefes são mais amigáveis e solidários com seus colegas do sexo masculino.

Para Hewlett et al. (2008), 52% das mulheres altamente qualificadas que trabalhavam para empresas voltadas para a área de ciência, engenharia e tecnologia, deixaram seus empregos por conta de ambientes de trabalho hostil. A cultura machista nestes ambientes exclui as mulheres dos espaços onde as tomadas de decisões acontecem e elas não são apenas fisicamente excluídas, mas muitas vezes são afastadas por comportamentos humilhantes ou degradantes.

A mídia, os indivíduos preconceituosos e a cultura em suas dimensões institucionais e organizacionais, são responsáveis por construir este ambiente como um domínio masculino onde as mulheres não devem entrar, exceto para ocupar cargos de *status* inferior (COLATRELLA, 2012). Estes padrões de comportamento contribuem para o fato de que mulheres relatam níveis mais baixos de satisfação com suas carreiras do que os homens, além de influenciar na probabilidade de que essas mulheres abandonem estes ambientes de trabalho (ASHCRAFT, 2016).

Nos últimos anos, houve uma quantidade crescente de pesquisas e discussões sobre a presença da mulher na área de tecnologia utilizando plataformas de comunicação relacionadas e não relacionadas a trabalho, como as redes sociais, apresentados nos estudos de Imtiaz et al (2019) e Ortu et al (2017) .

Segundo Boyd e Ellison (2010), redes sociais *online* são serviços baseados na web que permitem indivíduos criarem um perfil, sendo ele público ou privado. Através disso, os usuários podem se apresentar e se tornarem visíveis em suas redes, além de conectar-se uns com os outros, visualizando e explorando tanto a sua lista de conexões como as de outras pessoas. Com essas possibilidades, as redes sociais se transformaram na maior ferramenta

para comunicação entre pessoas e se tornaram uma plataforma para que pudessem expressar experiências e opiniões sobre diversos assuntos.

Uma dessas redes sociais é o Twitter, sendo uma das redes sociais mais populares mundialmente, possui mais de 271 milhões de usuários ativos por mês, que compartilham aproximadamente 500 milhões de tweets (mensagens de até 240 caracteres) por dia (TWITTER, 2014). O fluxo constante desses usuários gera uma enorme quantidade de dados, que tem despertado pesquisadores para coletar e utilizar estes dados para realizar análises para responder perguntas de suas pesquisas, como por exemplo, um padrão de comportamento de usuários dentro de uma plataforma.

No entanto, é inviável realizar a análise desses dados sem o auxílio de uma ferramenta, para solucionar este problema surgem os sistemas de análise de sentimentos. Análise de sentimentos é uma área que tem recebido atenção considerável recentemente. Segundo Liu (2012), a análise de sentimentos - também chamada de mineração de opinião - “é o campo de estudo que analisa opiniões, sentimentos, avaliações, atitudes e emoções para entidades como produtos, serviços, organizações, indivíduos, problemas, eventos, tópicos e seus atributos”.

A análise de sentimentos pode ser realizada em 3 diferentes níveis de granularidade: documentos, sentenças e aspectos. O nível que será utilizado no presente trabalho é o nível de aspectos. Este nível propõem identificar e extrair aspectos de um texto e palavras que exprimem uma opinião, e assim determinar a polaridade (positiva, negativa ou neutra) do aspecto relacionado a uma opinião.

A partir dessas possibilidades, é possível selecionar perfis em redes sociais de mulheres da área de tecnologia da informação e realizar uma análise sobre dados extraídos das conexões entre os usuários. A partir dessa análise pode-se criar possibilidades para possivelmente identificar gargalos a respeito da presença da mulher na área de tecnologia da informação, fornecer incentivos à diversidade de gênero e um ambiente de trabalho agradável.

O objetivo do presente trabalho é utilizar técnicas computacionais para extrair aspectos de tweets na língua portuguesa publicados por um conjunto de mulheres selecionadas que atuam na área de tecnologia da informação. Seus objetivos específicos são:

- Selecionar perfis de mulheres que atuam na área de tecnologia da informação;

- Coletar *tweets* utilizando a API disponibilizada pelo Twitter;
- Filtrar os *tweets* relacionados ao contexto de tecnologia;
- Realizar um tratamento na base de *tweets* coletados;
- Extrair aspectos frequentes em toda a base de dados.

Este trabalho está estruturado da seguinte forma: A seção 2 apresenta o referencial teórico sobre os assuntos abordados no trabalho: análise de sentimentos, análise de sentimentos em pequenos textos, extração de aspectos e por fim, aprendizado supervisionado, semi-supervisionado e não supervisionado baseado em léxicos. Na seção 3 é apresentada a metodologia e os materiais utilizados no desenvolvimento do trabalho. A seção 4 apresenta os resultados obtidos e a forma como foi desenvolvido e por fim na seção 5 é disposto as considerações finais.

## 2 REFERENCIAL TEÓRICO

### 2.1 ANÁLISE DE SENTIMENTOS

De acordo com Hussein (2016), a análise de sentimentos é o uso de mineração de dados, aprendizagem de máquina, linguística, processo de linguagem natural e técnicas de análise de texto para automatizar a identificação e classificação de sentimentos a partir de textos escritos em língua natural. O objetivo é analisar textos e determinar a opinião de pessoas, sentimentos, avaliações, atitudes, em relação a entidades, que podem ser produtos, serviços, organizações, eventos, grupo de pessoas, tópicos e seus atributos (LIU, 2010).

Liu (2012) sistematiza a análise de sentimentos em conjunto de termos sendo eles:

- **Entidade:** É o alvo da análise. Possui um conjunto de componentes e um conjunto de atributos e cada componente pode ter seus subcomponentes e seus atributos;
- **Aspectos:** Componentes e atributos de uma entidade;
- **Polaridade:** É a orientação de uma opinião sobre um aspecto, indicando se é positivo, negativo ou neutro;
- **Autor:** É a pessoa ou organização que expressa a opinião;
- **Tempo:** É o momento específico que a opinião foi expressada.

Segundo Liu (2012), existem diferentes níveis de análise textual, distinguidas entre três níveis de granularidade: de documento, de sentença ou de aspecto.

- **Nível de documento:** Classifica se uma opinião expressa em um texto é positiva ou negativa sobre uma única entidade, não sendo aplicado a um texto que avalie mais de uma entidade;
- **Nível de sentença:** Classifica se as sentenças de um texto são positivas ou negativas, ou seja, o texto analisado é dividido em sentenças e a análise é feita sobre cada unidade textual extraída. Este nível é bastante utilizado quando um mesmo documento contém opiniões sobre várias entidades;

- **Nível de aspectos:** Classificam se as entidades e/ou aspectos de um texto são positivos ou negativos. Um documento pode conter opiniões sobre diversos aspectos de diversas entidades.

Dentro da análise de sentimentos, ainda há vários desafios quanto à fonte textual a ser escolhida para tratamento e qual nível de análise deverá ser utilizado, pois os sentimentos não expressados da mesma forma em textos jornalísticos, blogs, mensagens em redes sociais, comentários e em decorrência disso, a identificação da polaridade do sentimento pode ser alterada. (Balahur et al, 2010).

## 2.2 ANÁLISE DE SENTIMENTOS EM PEQUENOS TEXTOS

Plataformas como Facebook, Instagram, LinkedIn e Twitter se tornaram extremamente populares no cotidiano das pessoas, facilitando a comunicação e compartilhamento de informações dos seus usuários. Essa conexão entre usuários gera conteúdos variáveis, desde um simples “like” em um tweet, a publicações de notícias em blogs.

Um estudo publicado por Naskar (2020), investiga a mudança das características das opiniões das pessoas contra um evento no Twitter e como discussões/interações em massa influenciam a mudança de opinião de alguém contra um evento. É utilizado a emoção de um usuário refletida na postagem como matriz para indicar sua opinião em apoio/oposição a um evento social.

O Twitter se trata de um *micro-blogging* criado em 2006 e que é utilizado por milhões de pessoas em todo o mundo, sendo a rede social mais popular da Internet. A interface desta rede social permite que seus usuários escrevam mensagens, que são chamadas de *tweets*, de até 240 caracteres que podem ser lidas pelos seus seguidores e, nos perfis públicos, por qualquer usuário do Twitter. (AMARAL; SOUSA, 2010)

Segundo Java et al. (2007), existem três categorias principais de utilizadores do Twitter: fontes de informação (jornais, revistas, sites de notícia), amigos (família, colegas de trabalho) e pessoas que buscam por informações. Neste estudo, os autores também identificaram várias categorias de intenção de uso do Twitter: conversa, conversa diária, compartilhamento de informações e relatos de notícias.

### 2.2.1 - Desafio em tratar *tweets*

Há uma grande quantidade de pesquisas na área de análise de sentimentos, porém a maior parte concentra-se em textos maiores como notícias e resenhas. Plataformas de comunicação de textos curtos e informais são cenários de pesquisas desafiadoras, um exemplo de plataforma é o Twitter.

Go et al (2009) aborda um estudo com uma área de pesquisa específica para o Twitter, conhecida como *Twitter Sentiment Analysis* (TSA). Diferente de resenhas e notícias que representam pensamentos resumidos de autores e ocasionalmente escritos de maneira formal, os *tweets* tendem a lidar com dificuldades adicionais na análise de sentimentos comparados às fontes de dados tradicionalmente usadas. Tamanho do texto, variação na ortografia, esparsidade dos dados, definição de contexto e negação são alguns dos desafios listados abaixo (SILVA, 2016).

- **Tamanho do texto:** *Tweets* são geralmente curtos, pois o comprimento máximo da mensagem é de 240 caracteres.
  - **Variação da ortografia:** Em consequência da limitação de caracteres e a informalidade, a ortografia em *tweets* são variáveis, devido aos erros ortográficos, abreviações, repetições de letras e sílabas.
  - **Esparsidade dos dados:** A variação da ortografia pode gerar um impacto no desempenho na classificação do sentimento, causando uma esparsidade dos dados. Segundo os estudos realizados por (Saif et al, 2012), a principal causa é o intervalo de frequência de palavras, onde as palavras se repetem de 10 a 20 vezes. Outro estudo de Seif et al (2012), demonstra um método para reduzir a esparsidade dos dados que é através da remoção de palavras irrelevantes para produzir resultados mais precisos.
- Negação:** Detecção de palavras de negação na análise de sentimentos é importante para classificação do sentimento, podendo inverter a polaridade de um texto.
- Stop words:** São consideradas palavras que possuem baixo poder de classificação de sentimento (como, "a", "que", "e") e são filtradas na etapa de processamento do texto.
- **Símbolos especiais e *emojis*:** Emoticons (símbolos como ":)", ":("), URLs, hashtag (uso de uma palavra após uma cerquilha #), e *emojis* (evolução de pictogramas como emoticons) são muito utilizados em textos provenientes de redes sociais como o Twitter. Estes símbolos especiais podem trazer dificuldades a ferramentas que realizam o processamento de linguagem natural de textos.

- **Variação de tópico:** Por ser uma rede social mundial, os tópicos discutidos no Twitter são ilimitados e de grande variação, causando problemas para a análise de sentimentos pois diferentes palavras podem expressar sentimentos diferentes em determinados contextos onde são utilizadas. Um estudo realizado por Zheng e Han (2013), são divididos os tópicos entre pessoal e público, onde tópicos pessoais são discutidos por vários usuários e retratam fortemente os interesses pessoais de um usuário, enquanto tópicos públicos geralmente são interessados por toda a comunidade e podem retratar fortemente questões quentes da sociedade. Ainda neste estudo, é considerado o símbolo hashtag "#" como indicador de tópico.
- **Quantidade de dados:** Devido a limitação do número de caracteres de tweets, a quantidade de textos publicados são extremamente altas. Em 2014, durante o jogo entre Brasil e Alemanha na copa do mundo, o Twitter publicou em seu blog que a rede social acumulou mais de 672 milhões de *tweets*. Nessa proporção de dados, uma opinião sobre um determinado tópico é dificilmente extraída.
- **Estilo de linguagem:** Os usuários do Twitter, ou seja, o linguajar e o estilo de escrita também são variáveis, podendo ser estritamente formais a informais com uso de gírias e expressões idiomáticas (que mudam o significado conforme o contexto). Com a mudança de vocabulário, pode conduzir problemas a análise de sentimentos ao lidar com dados treinados e anotados com sua classificação prévia.

**Contexto multilingual:** Usuários de redes sociais como o Twitter podem utilizar diversas línguas em seus textos publicados, seja na mesma mensagem ou até mesmo na mesma frase.
- **Tokenização:** A tokenização de sentenças é a quebra de sequência de caracteres de um texto em palavras, podendo ser usado duas abordagens: divisão por espaços em brancos ou não segmentadas. Essa etapa é outro desafio na análise de sentimentos, em decorrência disso, Owoputi et al (2013) apresentaram um tokenizador específico para ser utilizado no Twitter, baseado em outro tokenizador proposto por Gimpel et al (2012).

### 2.3 EXTRAÇÃO DE ASPECTOS

Aspectos se referem aos atributos e componentes de entidades. Segundo Liu (2015), aspectos podem ser extraídos através da relação sintática de sentenças e ao encontrar um substantivo ou um sujeito nominal é possível identificar o aspecto ligado. Em Liu (2012) é

apresentado as 4 principais abordagens a serem realizadas na etapa de extração de aspectos: extração baseada em frequência, extração baseada em relação sintática, extração usando aprendizado supervisionado e extração usando modelos de tópicos.

A extração baseada em frequência, a abordagem utilizada neste trabalho, busca encontrar expressões representadas por substantivos e/ou sujeitos nominais em aspectos explícitos. Inicialmente, é identificado os substantivos e sujeitos nominais usando um etiquetador morfológico (POS tagging) que determina qual a classe gramatical de cada palavra de uma sentença e, em seguida, conta suas frequências em todo o conjunto de dados analisados. Essa abordagem funciona pois os possíveis aspectos que são expressos como substantivos e sujeitos nominais voltam a se repetir por meio de outros usuários que também estão relatando suas opiniões e experiências sobre os diferentes aspectos de uma entidade. (HU;LIU, 2004)

A abordagem por meio de relações sintáticas foi proposta no estudo de Hu & Liu (2004), onde observaram a relação entre o alvo e o sentimento e utilizaram palavras de sentimento para identificar aspectos pouco frequentes. Esse relacionamento é explorado para extrair entidades e aspectos, pois as palavras de sentimento são conhecidas por meio de recursos léxicos. Portanto, se alguma palavra de sentimento for não for conhecida, essa relação também pode ser usada para extraí-la.

A abordagem de aprendizado supervisionado trata a extração de aspectos como uma tarefa de classificação e exigem grandes corpus rotulados para treinamento. No estudo de Yu et al. (2011) foi utilizado modelos baseados em SVM (*Support Vector Machine*) para classificar os aspectos em um relato. Cada instância representa um termo que é classificado como aspecto e não-aspecto. Todos os aspectos classificados são extraídos e representados através de sumários, além disso também é identificado expressões e sinônimos de aspectos frequentes. Porém, é apresentada uma limitação na utilização dessa abordagem, pois a identificação de aspectos só ocorre a partir de uma lista limitada de substantivos frequentes.

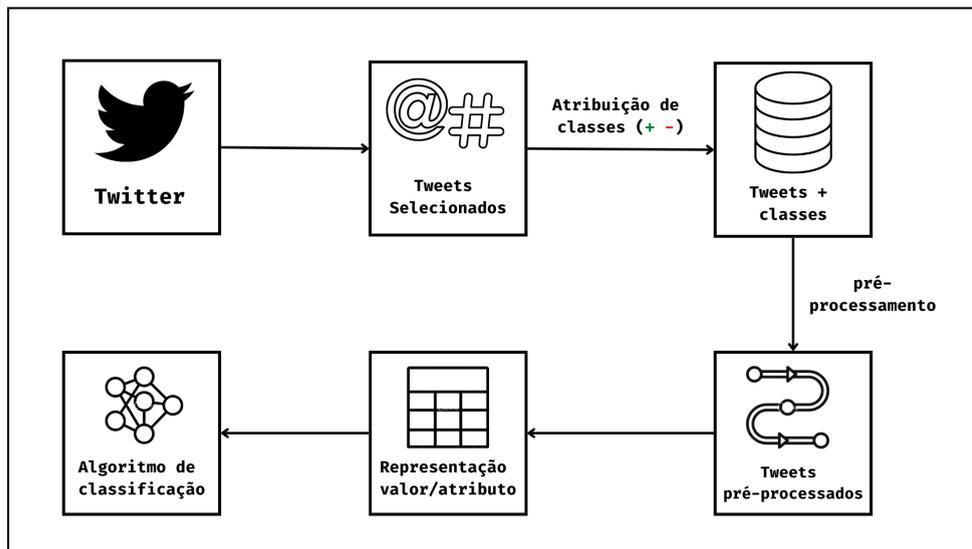
Por fim, a quarta abordagem, baseada em tópicos, utiliza-se de modelos estatísticos para identificar tópicos em grandes *corpus*. Esses tópicos correspondem a diferentes aspectos, palavras opinativas, ou locais citados no texto. Os modelos de extração de tópicos probabilísticos esperam que os documentos analisados consistam em uma mistura de tópicos, e que um tópico seja uma distribuição probabilística sobre as palavras do corpo. (LIU, 2015)

## 2.4 APRENDIZADO SUPERVISIONADO

Essa abordagem faz uso de aprendizado de máquina, composto por algoritmos que associam uma entrada com uma saída com base em dados pré-rotulados, onde o algoritmo possui pares de entrada e saída, geralmente em forma de vetores. Em cada saída é atribuído a ela um valor numérico ou uma classe, assim é previsto um rótulo de saída conforme a entrada informada no par. (FONTANA, 2020).

Em uma abordagem de aprendizado supervisionado de máquina exige dois conjuntos de documentos: um para treinamento e outro de teste. O de treinamento é composto por instâncias que são formadas por vários atributos e estão previamente rotulados com suas polaridades nas classes de interesse, geralmente realizado manualmente por um especialista de domínio. Após a obtenção de dados rotulados, é necessário classificar os dados para a distinção e transformá-los em um modelo de representação. Uma representação geralmente utilizada é chamada de *bag of words*, onde cada texto é classificado por 0's e 1's, onde 1 indica a presença de uma palavra, e 0 a ausência. (SILVA, 2016; BENEVENUTO et al, 2015)

**Figura 1:** Fluxo de aprendizagem supervisionado de máquina



**Fonte:** Adaptado de Silva (2016)

O algoritmo assim aprende um modelo de classificação que pode ser posteriormente aplicado aos documentos que não estão rotulados (conjunto de testes). SVM (*Support Vector*

*Machine*), *Naïve Bayes* (NB) e Máxima Entropia (MAXENT) são exemplos de algoritmos aplicados nesta abordagem e sendo métodos propostos para lidar com TSA.

## 2.5 APRENDIZADO SEMI-SUPERVISIONADO

A abordagem de aprendizado semi-supervisionado se baseia no aprendizado supervisionado, porém, além de presumir entradas de dados rotulados, presume-se a entrada de dados não rotulados. O objetivo é utilizar as informações rotuladas para extrair o problema e guiar o aprendizado a partir dos dados não rotulados. No contexto de análise de sentimentos para o Twitter (TSA), essa abordagem pode ser considerada por conta da dificuldade de se encontrar coleções de tweets em língua portuguesa pré-rotulados. (BRUCE, 2001).

No trabalho de Silva (2016) é identificado três métodos de abordagens para o aprendizado semi-supervisionado em TSA, sendo eles: baseado em grafos, baseado em *self-training* e *co-training* e baseado em tópicos. Em método baseado em grafos, dados de entrada (*tweets*) são propagados para os dados não rotulados, onde o processo de propagação de rótulos necessita do cálculo das similaridades entre as instâncias dos dados.

Em método baseado em *self-training* e *co-training* é realizado por iterações, onde em cada iteração, uma certa quantidade de instâncias não rotuladas é rotulada pela função de decisão que é aprendida e incorporada aos dados de treinamento. A partir de suas próprias previsões e dos dados rotulados, o modelo de classificação é treinado novamente para a próxima iteração.

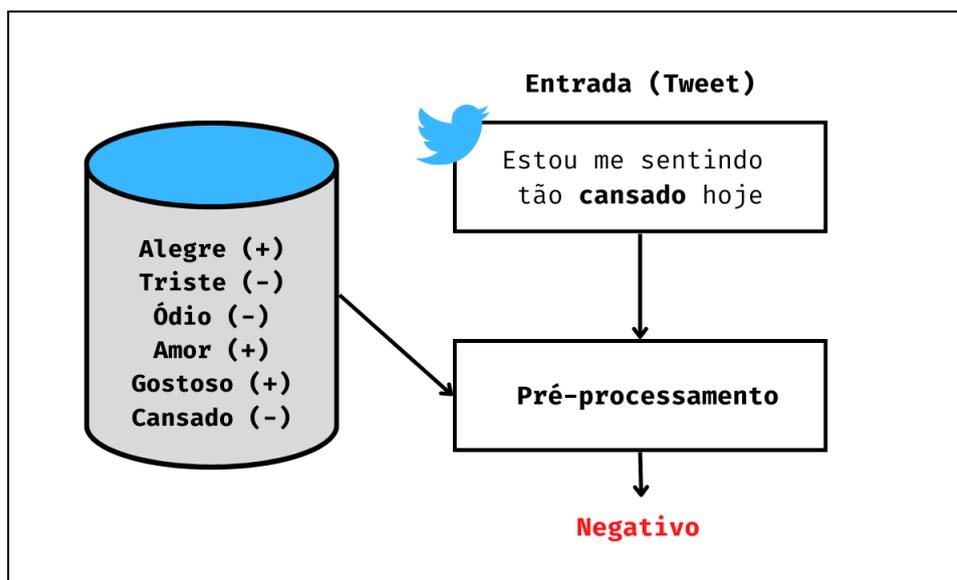
No método baseado em tópicos é proposto a construção de um modelo de tópico em tweets rotulados para que modelos de sentimento específicos possam ser induzidos em cada grupo encontrado. Para cada tópico, o sentimento é derivado de acordo com uma distribuição de palavras de opinião visando para construir um léxico de sentimentos.

## 2.6 APRENDIZADO NÃO SUPERVISIONADO BASEADO EM LÉXICO

Essa abordagem, diferente do aprendizado supervisionado, não utiliza dados pré-rotulados e documentos para a criação de um modelo de aprendizagem e sim faz o uso de léxicos (dicionários) de sentimentos. Esse léxico de sentimentos são dicionários de palavras, ao invés de possuir como conteúdo o significado de cada palavra, possui em seu lugar um significado quantitativo sendo um número entre -1 a 1, onde -1 é o valor sentimental mais negativo e 1 o valor mais positivo. Abordagens léxicas assumem que palavras possuem uma

polaridade prévia, ou seja, uma orientação semântica independente de contexto e que podem ser expressadas com um valor numérico ou classe. (TABOADA, 2011).

**Figura 2:** Fluxo de aprendizagem não supervisionado baseado em léxicos



**Fonte:** Adaptado de Benevenuto (2015)

A figura 2 representa o fluxo de uma análise de sentimentos baseada em léxico. O primeiro passo da classificação é a entrada de um texto, no caso do exemplo, um *tweet*. O segundo passo é realizar o pré-processamento para realizar um tratamento no texto e por fim uma pesquisa no dicionário de termos léxicos. Ao fim do processo, o método é capaz de definir qual a polaridade ou sentimento na sentença de texto na entrada. (BENEVENUTO, 2015).

A construção desses dicionários podem ser divididos entre duas classificações: orientação semântica e intensidade de sentimento. Léxicos de orientação semântica são baseados em polaridade, ou seja, classifica as palavras por positivas e negativas. LIWC, SentiStrength e WordNet são exemplos de dicionários para lidar com análises de orientação semântica. Léxicos de intensidade de sentimento vão além de determinar apenas a polaridade binária (positiva e negativa), mas também a força do sentimento expresso no texto, trazendo a possibilidade de reconhecer mudanças na intensidade do sentimento ao longo do tempo. (HUTTO; GILBERT, 2014).

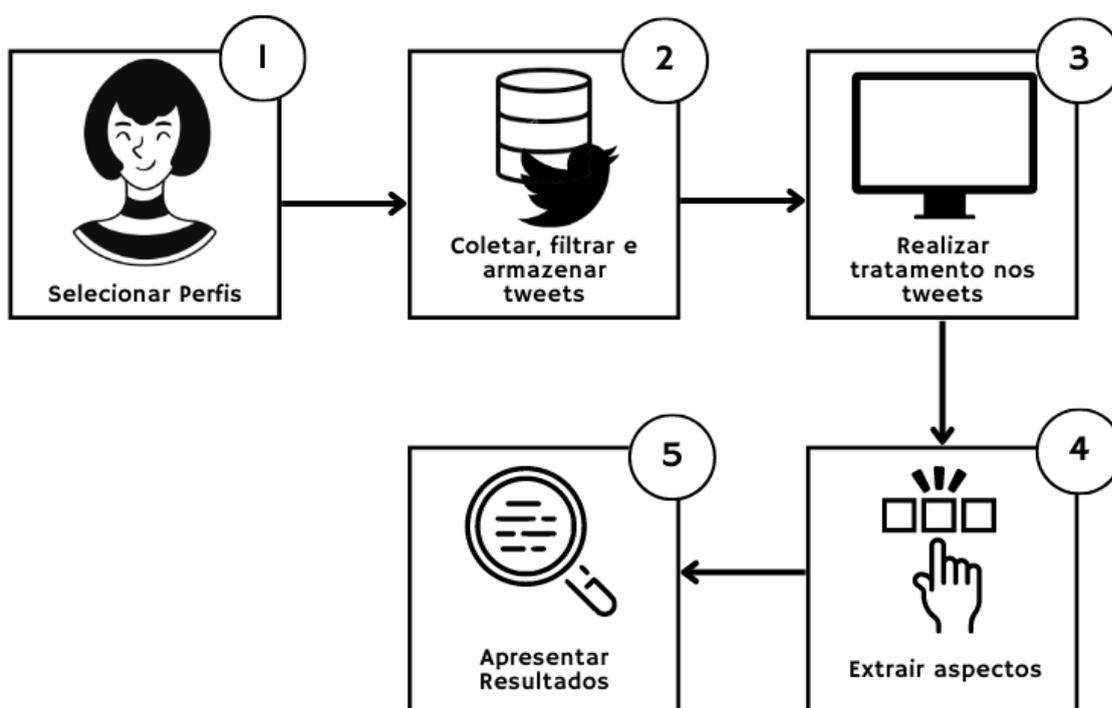


### 3 METODOLOGIA

#### 3.1 MATERIAIS

Os materiais que foram utilizados para o desenvolvimento do trabalho proposto são: A linguagem de programação Python, a API do Twitter, o banco de dados PostgreSQL, a biblioteca de processamento de linguagem Natural NLTK, e a biblioteca `spaCy`. Na figura abaixo é apresentado o fluxo de execução do trabalho e o detalhamento das etapas que foram realizadas no desenvolvimento do trabalho.

Figura 3: Fluxo de Trabalho



Na etapa 1 foi feita a seleção manual de perfis de mulheres que atuam na área de tecnologia da informação para a extração dos *tweets* e seus comentários. Os critérios para a seleção dos perfis foram: mulheres que atuam na área da tecnologia de alguma forma, por exemplo, trabalham na área, criadoras de conteúdo, participantes de comunidade de código aberto. E também possuem ao menos 5 mil seguidores, para garantir uma grande troca de *tweets* e assim também tendo uma grande quantidade de dados a serem analisados. Foram selecionadas 15 mulheres baseado nos critérios acima, sendo todas ativas recorrentemente na rede social.

Na etapa 2 foram coletados 19.468 *tweets* realizados durante um período de 6 meses por cada perfil e a partir deles foram filtrados os comentários. Posteriormente, esses *tweets* foram armazenados no banco antes da etapa de tratamento de dados para manter a base de dados original. Para esta coleta e filtragem dos *tweets* foi utilizado a linguagem Python, que é uma linguagem de programação de alto-nível, dinâmica, orientada a objetos e funcional, mantida pela *Python Software Foundation* (PSF). Inclui estruturas de alto nível como listas, dicionários, complexos, além de módulos prontos para uso, frameworks de terceiros como Django, Flask, que fornecem ferramentas para desenvolvimento de projetos. É uma linguagem multiparadigma, que suporta programação modular e funcional, além de orientada a objetos. (PYTHON, 2022). E o TwitterAPI, que é uma API (*Application Programming Interface*) desenvolvida pela rede social Twitter, que permite que desenvolvedores acessem as principais funcionalidades do Twitter utilizando métodos padronizados como: postar *tweets*, pesquisar *tweets*, filtrar *tweets*, acessar seguidores de um usuário, (TWITTER, 2022).

Para o armazenamento dos *tweets* foi utilizado PostgreSQL, um sistema gerenciador de banco de dados relacional de código aberto criado a 30 anos pelo *PostgreSQL Global Development Group* (POSTGRESQL GLOBAL DEVELOPMENT GROUP, 2022). Nesse banco de dados será armazenado todos os dados coletados (*tweets*) e o resultado do pré-processamento dos dados e da análise de sentimentos.

Na etapa 3 foi realizada a etapa tratamento na base dos *tweets* coletados, foram aplicados a normalização (remoção de símbolos especiais pontuação, emoticons, emojis e tags), remoção de *stop-words*, lematização, correção ortográfica, e *pos-tagging* dos textos dos *tweets* extraídos. Após o tratamento, os *tweets* foram armazenados no banco de dados. Para essa etapa, foi utilizado a biblioteca NLTK (*Natural Language Toolkit*) ou Kit de ferramentas de linguagem natural, uma biblioteca Python de código aberto para processamento de texto desenvolvido em 2001 pela Universidade de Pensilvânia. A biblioteca contém um conjunto de módulos para classificação, tokenização, lematização, marcação, análise e raciocínio semântico. (NLTK, 2022).

Na etapa 4 foi desenvolvida a identificação dos aspectos dos *tweets* coletados da etapa anterior. A partir da etiquetagem morfológica realizada na etapa de *pos-tagging*, foi identificado os aspectos de cada sentença. Cada aspecto foi armazenado em uma planilha e essa planilha foi utilizada para verificar a frequência de cada aspecto selecionado. Para isso foi utilizado a biblioteca *spaCy*, uma biblioteca Python de código aberto que também traz

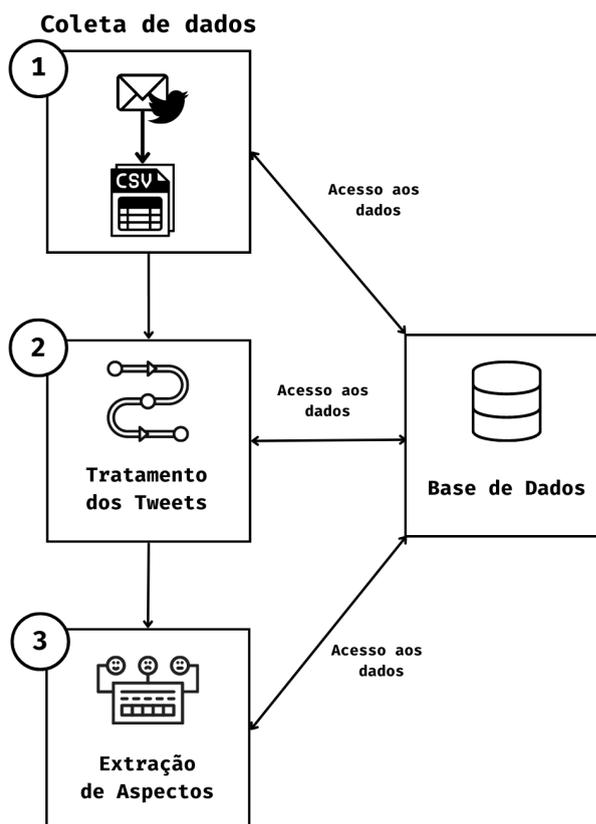
ferramentas para o processo de linguagem natural. Com o suporte para 72 idiomas e 80 pipelines treinados para análise de dependência, pos-tagging, tokenização, análise morfológica entre outros. (SPACY, 2022)

Na etapa 5 foi gerado e apresentado o relatório dos resultados da extração de 2130 aspectos coletados e o cálculo da frequência dos mesmos em toda a base de dados através de gráficos e tabelas.

## 4 RESULTADOS E DISCUSSÕES

Essa seção apresenta os resultados obtidos no desenvolvimento do trabalho e as discussões levantadas referentes aos resultados. Esta seção segue com a apresentação inicial das etapas para extração de aspectos e posteriormente os módulos desenvolvidos: coleta de dados, tratamento dos dados e identificação de aspectos como demonstrado na figura abaixo.

**Figura 4:** Arquitetura das etapas de extração de aspectos

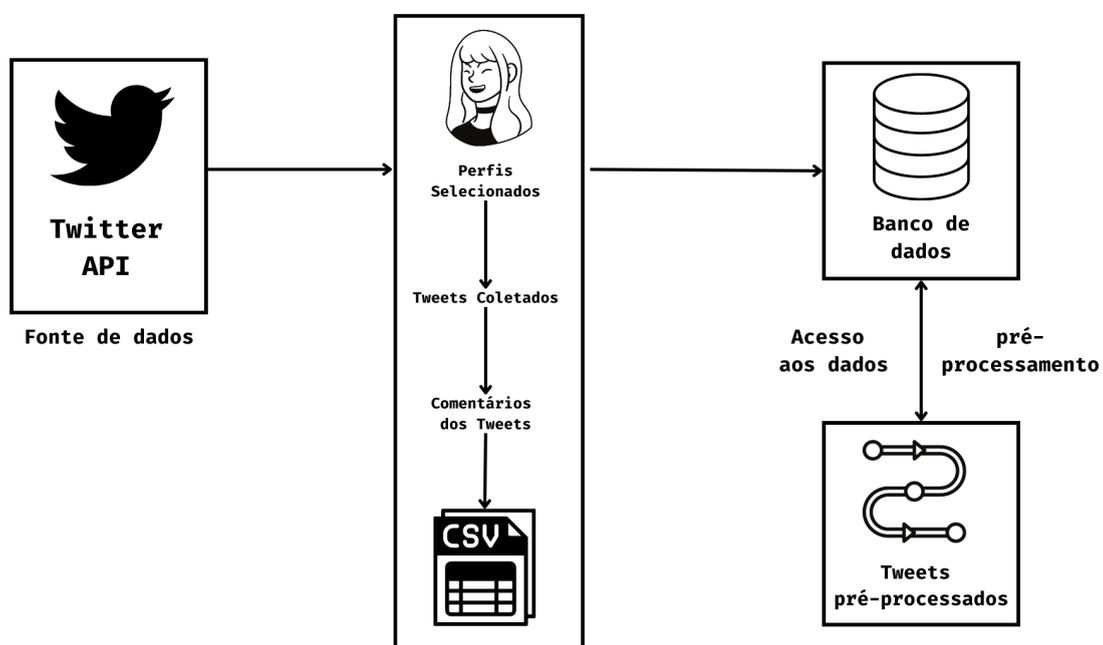


1. **Coleta de dados:** Etapa onde é feito a extração dos *tweets* e seus comentários de cada perfil selecionado para coleta de dados e a partir dessa coleta é gerado um arquivo CSV contendo todas as informações dos comentários realizados.
2. **Tratamento dos Tweets:** Realiza a leitura do arquivo CSV gerado na etapa anterior e inicia a o processamento de texto nos *tweets* para limpeza da base de dados,
3. **Extração de Aspectos:** Etapa que busca todos os tweets pré-processados na etapa anterior de tratamento de dados e realiza a identificação de aspectos e sua frequência na base de dados coletados.

## 4.1 COLETA DE DADOS

Para a coleta dos tweets foi desenvolvido um módulo em Python, utilizando uma biblioteca chamada `Tweepy` para o acesso à API (*Application Programming Interface*) do Twitter. A `Tweepy` permite um fácil acesso a todos os métodos da API como listar *tweets*, listar dados de um usuário e listar comentários de um *tweet*. A execução dessa etapa foi dividida entre três sub-etapas, sendo a primeira etapa a busca dos *tweets* realizados por pelos 15 perfis selecionados, filtrados por um período de 1 ano.

**Figura 5:** Arquitetura do módulo de coleta de dados

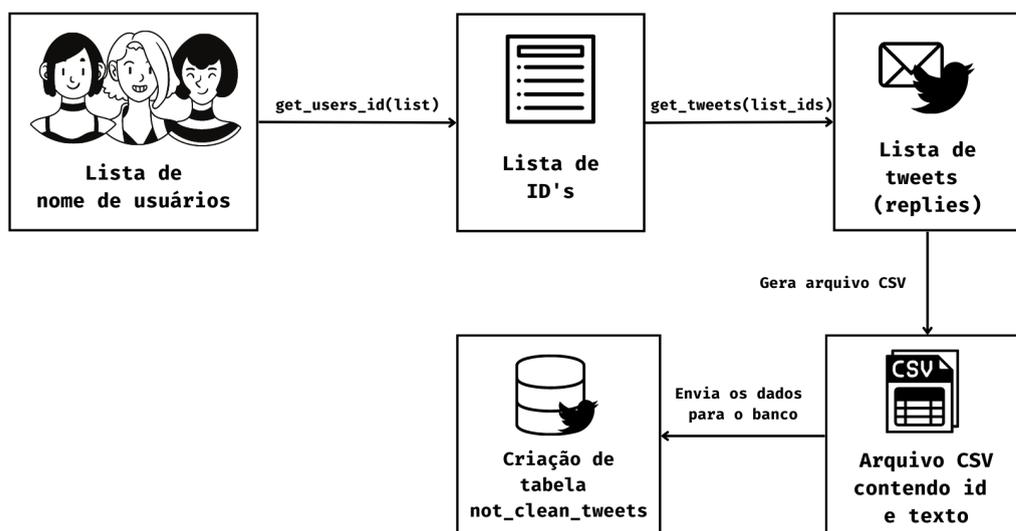


Foram coletados 19.468 *tweets* e a partir deles foi gerado um arquivo CSV ao final da coleta de dados e esses dados foram armazenados no banco de dados POSTGRESQL. No arquivo CSV contém o texto de cada tweet realizado e seu respectivo ID na rede social Twitter. Depois, na terceira etapa, é realizada a limpeza dos dados processados e enviados para o banco de dados em outra tabela para manter-se os dados originais e os dados tratados. As seções a seguir apresentarão com mais detalhes a realização da extração e tratamento dos dados.

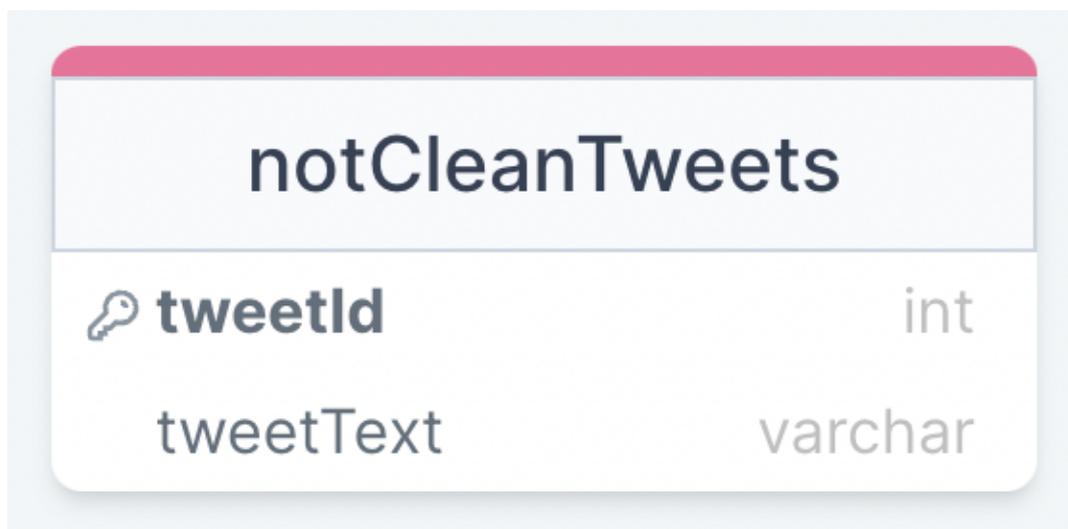
Para a extração dos tweets inicialmente foi definido o nome de usuários dos perfis selecionados, para isso foi criado um método chamado `get_users_id`. Esse método recebe como parâmetro uma lista com o nome de usuário de cada perfil na plataforma e tem como

função fazer uma requisição HTTP para API do Twitter para buscar o id correspondente ao nome de usuário e armazenar em outra lista para serem utilizados na busca dos *tweets*, realizados no método *get\_tweets*.

Figura 6: Fluxograma da extração dos *tweets*



Na figura acima é apresentado o método *get\_tweets*, que tem como função buscar todos os *tweets* de cada perfil selecionado durante um período de 6 meses, sem incluir *retweets* (a republicação de um *tweet* realizado por outro usuário). O método faz outra requisição HTTP para a API, que retorna uma lista contendo todos os dados contidos em um *tweet*, especialmente os *replies* (respostas) e o id de cada *reply* (resposta) que serão utilizados no projeto. A partir dessa lista foi criado um arquivo CSV, armazenando o id e o texto, onde esse arquivo é carregado para o banco de dados e é criada uma tabela *not\_clean\_tweets*, com as colunas *TweetId* e *TweetText*.

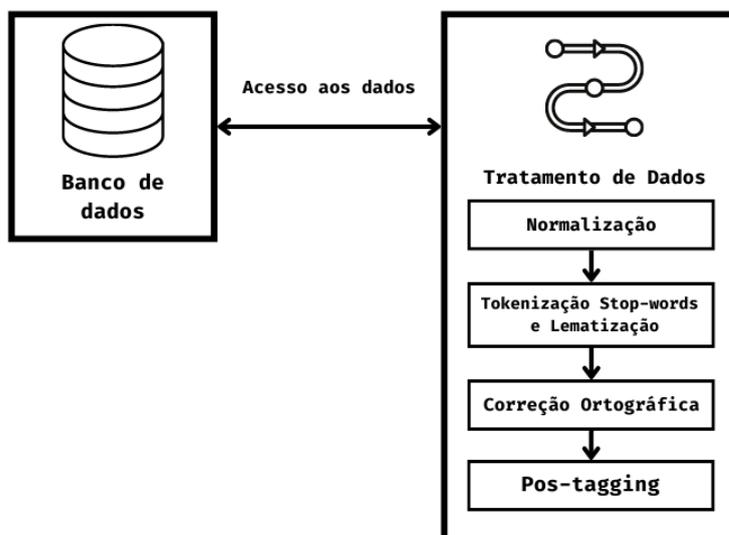
**Figura 7:** *Schema* do banco de dados

Na figura 7 é representado o *schema* do banco de dados criado no processo de extração de *tweets*. Nessa tabela são armazenados os dados que foram usados no módulo de tratamento de *tweets*, onde o id (número de identificação único do tweet é a chave primária da tabela (identificador único de registro em uma tabela no banco de dados). Antes de cada linha do arquivo CSV ser inserida na tabela, é verificado se o id e texto estão armazenados para evitar duplicidade dos dados.

#### 4.2 - TRATAMENTO DE TWEETS

O módulo de tratamento dos dados coletados é dividido em 4 partes correspondentes: normalização, tokenização e correção ortográfica, e *pos-tagging* conforme apresentado no fluxograma abaixo.

**Figura 8:** Fluxograma de tratamento dos *tweets*



Como representado na figura acima, na etapa de tratamento de dados foi criado um método chamado *clean\_tweets* que recebe como parâmetro todos os *tweets* coletados no primeiro módulo que estão salvos no banco de dados. Neste método é realizada a normalização dos dados inicialmente. Nas seções abaixo serão apresentados com detalhes o que cada etapa do tratamento foi feito.

A normalização é a primeira etapa do módulo de tratamento de *tweets*. Essa etapa realiza a padronização linguística necessária para lidar com as peculiaridades de uma base de dados como *tweets*. URLs, nome de usuários, hashtags são exemplos de peculiaridades. O tratamento desses dados é uma etapa crucial para o desempenho da classificação que será realizada no módulo de análise de sentimentos. Na tabela abaixo é apresentado um exemplo desse tratamento utilizando um tweet extraído na etapa de coleta de dados.

Texto Original	Texto Normalizado
<p><b>@evento</b> Você é mulher e quer atuar no projeto? Se inscreva aqui até dia <b>02/09</b>: <a href="https://t.co/434444">https://t.co/434444</a>. Compartilhe o projeto, deixe sua estrelinha no repositório para receber os updates. E quando estiver pronto, vamos gostar <b>muito</b> do seu feedback! :) <b>#evento</b></p>	<p>you are a woman and want to work on the project? sign up here until <b>02/09</b>: <a href="https://t.co/434444">https://t.co/434444</a>. share the project, leave your star in the repository to receive updates and when you are ready we will love your feedback <b>very</b></p>

**Tabela 1:** Exemplo de normalização

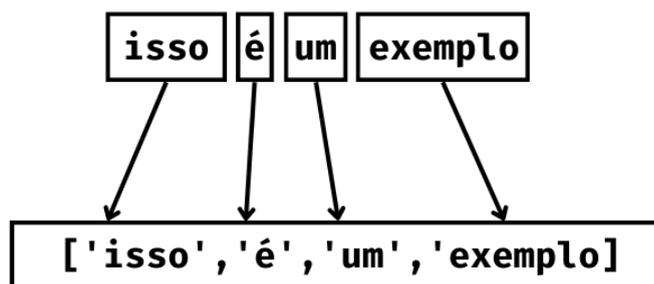
Todos textos que estão destacados na tabela acima foram removidos ou substituídos, conforme demonstrado na coluna à direita, e outros dados que também foram normalizados durante essa etapa estão indicados abaixo:

- Remoção de *hashtags*. Exemplo: remove #evento
- Remoção de *links*. Exemplo: remove "https://t.co/434444"
- Remoção de nome de usuário. Exemplo: remove a marcação de algum usuário @evento
- Remoção de caracteres especiais. Exemplo: remove caracteres especiais como \$, \*, e pontuação.
- Substituição de emoticons. Exemplo: substitui o emoticon de tristeza ":" por triste.
- Remoção de letras duplicadas. Exemplo: remove as letras duplicadas da palavra "muitoooo", se tornando "muito".

A lista completa de todas as remoções e substituições com suas expressões regulares utilizadas se encontram na seção APÊNDICE A.

A execução da segunda etapa do módulo de tratamento de *tweets* consiste em 3 sub-etapas, sendo elas: tokenização de palavras, remoção de *stop-words*, e *stemming* de palavras. A primeira sub-etapa é a tokenização, que consiste no processo de quebrar uma sentença em unidades simples de *tokens*. Palavras, sentenças e caracteres podem ser consideradas tokens. Para a realização da tokenização dos tweets, foi utilizado o módulo da biblioteca NLTK *word\_tokenization* e foi escolhida a tokenização por palavras.

**Figura 9:** Exemplo de tokenização.



Fonte: Autor

Na figura 9 é apresentado um exemplo de como é realizado na tokenização. A tokenização é utilizada para facilitar a interpretação de cada palavra na sentença para a realização das próximas etapas. O tweet é passado pela função quebrando o texto em tokens e armazenados em uma lista. Essa lista será utilizada para as etapas de remoção de *stop-words* e lematização.

*Stop-words* são palavras que atribuem somente significado sintático a uma sentença, exemplo dessas palavras são: *as, e, os, com, ou, para*. Essas palavras não adicionam significado a uma frase e é necessário removê-las para a diminuição do conjunto de dados e melhorar o desempenho de classificação que será realizado na extração de aspectos, um exemplo de um dos *tweets* coletados é demonstrado na tabela 2 abaixo.

Texto	Texto com remoção de stop-words
você é mulher e quer atuar no projeto se inscreva aqui compartilhe o projeto deixe sua estrelinha no repositório para receber updates e quando estiver pronto vamos gostar muito do seu feedback <b>feliz</b>	mulher quer atuar projeto inscreva aqui compartilhe projeto deixe estrelinha repositório receber updates pronto vamos gostar feedback feliz

**Tabela 2:** Exemplo de remoção de stop-words a partir de um *tweet* extraído na primeira etapa

A tabela 2 mostra o resultado da remoção de *stop-words* do mesmo exemplo mostrado anteriormente na etapa de normalização. Foi utilizado o *corpus stopwords* disponibilizado pela biblioteca NLTK em português, que recebe como parâmetro cada *token* da listagem realizada na etapa de tokenização e verifica se é uma *stop-word* ou não.

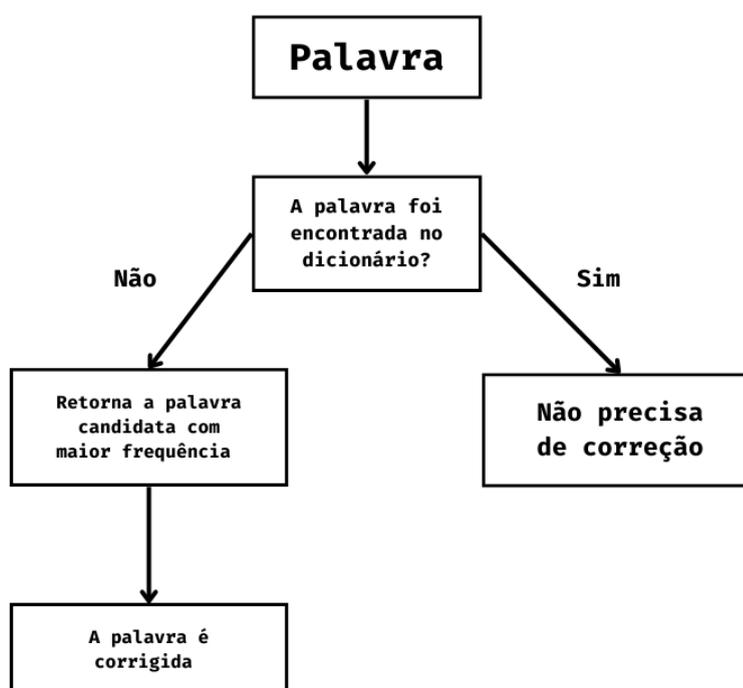
A lematização foi a última sub-etapa a ser realizada no módulo de normalização, que consiste em reduzir cada palavra da lista de palavras tokenizadas em sua palavra raiz, fazendo com que seja evitada a redundância de palavras. Por exemplo, os termos comer, comido, comendo tem como raiz a palavra come.

<b>Frase</b>	<b>Frase após a lematização</b>
mulher quer atuar projeto inscreva aqui compartilhe projeto deixe estrelinha repositório receber updates pronto vamos gostar feedback feliz	mulher quer atuar projeto inscreva aqui compartilh projeto deix estrelinha repositório receb updat pronto vamo gostar feedback feliz

**Tabela 3:** Exemplo de lematização a partir de um *tweet* extraído na primeira etapa

A terceira etapa do tratamento de tweets é a correção ortográfica. Essa etapa tem como objetivo corrigir a grafia de todas as palavras *tokenizadas* coletadas e processadas nas etapas anteriores. Para a realização dessa etapa foi utilizada a biblioteca `pyspellchecker`, construída em base do corretor ortográfico construído por Norvig (2016).

**Figura 10:** Exemplo do algoritmo de correção ortográfica.



Fonte: Autor

O algoritmo recebe como parâmetro uma palavra e ela é consultada no dicionário de palavras conhecidas, se a palavra estiver listada, não é corrigida. Caso a palavra não for encontrada, é feita uma pesquisa no dicionário e a palavra é corrigida pela palavra candidata com maior frequência. Um exemplo de entrada e saída do algoritmo está representado na tabela abaixo:

<b>Palavra Incorreta</b>	<b>Palavra Corrigida</b>
repositorio	repositório
virtualizao	virtualização
felis	feliz
proceso	processo

**Tabela 4:** Exemplo de correção ortográfica

O pos-tagging é a última etapa do tratamento de *tweets*. Essa etapa é responsável por identificar cada palavra em uma sentença, sendo um etiquetador morfológico que identifica a classe gramatical de cada *token* dentro agrupamento de classes chamada de PoS (Part-of-Speech), por exemplo: verbo, advérbio, adjetivos, substantivos. O conjunto dessas palavras montam o que é chamado de *tagset*. (JURAFSKY; MARTIN, 2000). O objetivo é reduzir a ambiguidade dentro do processo de tratamento de dados.

<b>Tag</b>	<b>Descrição</b>
NOUN	Substantivo
ADJ	Adjetivo
ADV	Advérbio
AUX	Auxiliar
N	Nome
ADP	Preposição
PR	Pronome
VERB	Verbo
SCONJ	Conjunção Subordinativa

Tag	Descrição
CONJ	Conjunção

**Tabela 5:** Exemplo de *tagset*

Na tabela acima é representado as tags utilizadas nessa etapa e sua respectiva classe gramatical dentro do agrupamento de classes PoS universal. Usando *tagsets*, as palavras serão etiquetadas como: verbos por VB, artigos por AR, adjetivos por AD e pronomes como PR, como exemplificado na figura 11 acima.

**Figura 11:** Exemplo de sentença etiquetada a partir de um *tweet* coletado.

PR            N            ADV            ADP            N            SCONJ            VERB            N

**Me motivo muito pela ideia de capacitar pessoas**

SCONJ            VERB                            N            ADV            ADP            N            ADJ

**para desenvolver tecnologia através do conhecimento coletivo**

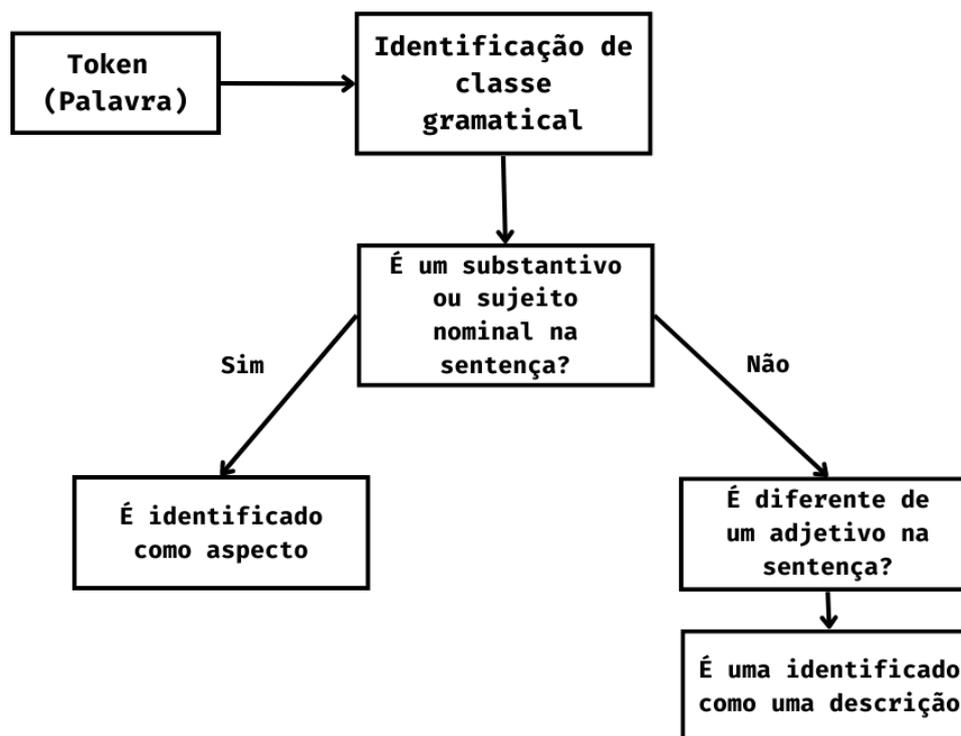
**Fonte:** Autor

Para o algoritmo para identificação de cada *tagset* foi utilizado a biblioteca spaCy, que possui um *corpus* treinado e classificado com mais de 500 mil palavras etiquetadas em português extraídos de corpos de texto de notícias, artigos e mídia. Foi criada uma função que recebe cada *token* (extraídos da etapa anterior) como parâmetro e consulta o *corpus* treinado e etiquetando cada token com sua respectiva *tagset*.

### 4.3 EXTRAÇÃO DE ASPECTOS

Essa etapa se concentra na extração de aspectos sobre cada *tweet* coletado. O processo de identificação de aspectos é realizado em cada *tweet* pré-processado na etapa de tratamento de *tweets*. O algoritmo percorre a cada *token* de um *tweet* identificando em qual classe gramatical o *token* pertence dentro da sentença. Caso seja identificado como um substantivo ou um sujeito nominal, a palavra é classificada como um aspecto. Na figura abaixo é exemplificado a identificação de um aspecto.

**Figura 12:** Exemplo de fluxo do algoritmo de extração de aspectos.



Fonte: Autor

Quando o aspecto é identificado, ele é adicionado a uma lista de aspectos que é criado um arquivo csv que será utilizado na próxima etapa responsável por definir a frequência dos aspectos. Para identificar a frequência de cada aspecto selecionado, foi criado uma função que percorre o arquivo csv com a lista de todos os aspectos coletados, verificando quantas vezes o mesmo aspecto esteve presente na listagem. Por fim, o algoritmo retorna os aspectos e quantidade em ordem de maior número de frequência.

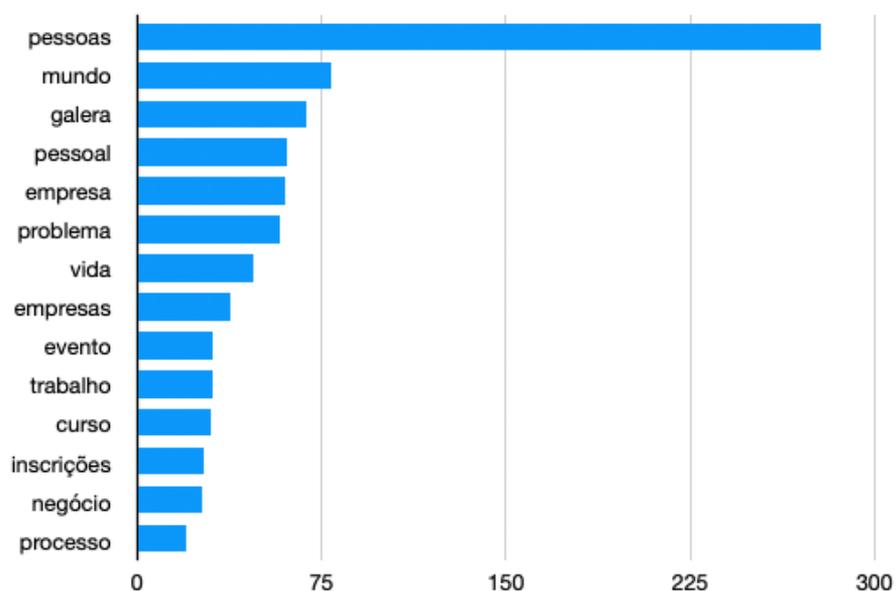
Foram extraídos 2130 aspectos com suas descrições e frequências em todos os 19.468 tweets coletados. A organização do resultado foi realizada em duas planilhas. A primeira planilha contém todos os aspectos coletados com suas respectivas descrições, um exemplo de alguns extraídos será apresentado na figura abaixo.

**Figura 13:** Tabela com alguns aspectos selecionados

Aspecto	Descrição
equipe	incrível
produto	simples
serviço	bom
time	responsável
hotel	diferente
álbum	sensacional
teste	urgente
evento	maravilhoso

Fonte: Autor

A segunda planilha extraída é o resultado de todos os aspectos e sua respectiva frequência em todos os *tweets* coletados. A lista completa com os aspectos e sua respectiva frequência se encontram na seção APÊNDICE B. Na figura abaixo é apresentado um gráfico com alguns aspectos mais frequentes e o valor de vezes que foram mencionados. Observa-se no gráfico que muitas palavras frequentes possuem relação entre si e com o contexto buscado no trabalho, que é a área da tecnologia. Palavras como: evento, empresa, trabalho, curso, negócio, processo, são palavras presentes no dia a dia na comunicação entre pessoas da área.

**Figura 14:** Gráfico com aspectos selecionados e sua frequência

**Fonte:** Autor

Porém, a utilização do Twitter trouxe dificuldade de atender todos os respectivos pontos iniciais para realizar a análise de sentimento de algum texto, pois seu ponto principal é a identificação de aspectos. O contexto abordado não possui aspectos e entidades definidos com suas respectivas polaridades, a falta desses dados podem tornar difícil a realização de uma análise de sentimentos relevantes e com alta precisão.

## 5 CONSIDERAÇÕES FINAIS

O presente trabalho teve como objetivo realizar a extração de aspectos em comentários de *tweets* na língua portuguesa de mulheres presentes na área da tecnologia. O trabalho foi dividido em coleta de dados, tratamento de *tweets* e identificação e extração de aspectos.

Na coleta de dados, a utilização da biblioteca `tweepy` foi uma boa alternativa para a extração dos *tweets*, sendo possível filtrar por data, por quantidade de *tweets* e por cada perfil selecionado. Porém, houve uma limitação de acesso a API do Twitter por conta da divisão de acessos que a plataforma restringe e neste trabalho não foi possível acessar o acesso de maior nível, o que causou uma dificuldade para a busca direta dos comentários de cada *tweet*.

Na etapa de tratamento de *tweets*, a biblioteca `NLTK` foi de maior importância para todas as sub-etapas realizadas. O módulo de tokenização de palavras, remoção de *stop-words* e lematização para a língua portuguesa foi preciso. A etapa de *pos-tagging* utilizando a biblioteca `spaCy` desempenhou um ótimo papel para realizar a etiquetagem morfológica nos *tweets*, porém houve uma dificuldade com a presença de palavras em inglês dentro das sentenças escritas na língua portuguesa.

Na etapa de extração de aspectos, embora não tenha sido possível continuar para a etapa de análise de sentimentos e não foram encontrados aspectos e entidades bem definidos, estes resultados podem ser utilizados como ponto de partida para trabalhos futuros.

Embora o trabalho tenha sido desenvolvido para o contexto da presença de mulheres na área da tecnologia, as etapas de tratamento de *tweets* e identificação de aspectos foram construídas de forma independente, sendo possível utilizar estas etapas em outros contextos, porém apenas em textos escritos na língua portuguesa.

## 6 REFERÊNCIAS

AMARAL, Inês; SOUSA, Helena. **Redes Sociais no Twitter: A Emergência de uma Nova Sociabilidade num Novo Ecossistema de Comunicação?** Interações: Sociedade e as novas modernidades, v. 10, n. 19, 31 out. 2010.

ASHCRAFT, Catherine; MCLAIN, Brad; EGER, Elizabeth. **Women in tech: The facts.** Colorado, CO, USA: National Center for Women & Technology (NCWIT), 2016.

BALAHUR, Alexandra; STEINBERGER, Ralf; KABADJOV, Mijail; ZAVARELLA, Vanni; GOOT, Erik van Der; HALKIA, Matina; POULIQUEN, Bruno; BELYAEVA, Jenya. **Sentiment Analysis in the News**, 2010.

BENEVENUTO, Fabrício; RIBEIRO, Filipe; ARAÚJO, Matheus. **Métodos para Análise de Sentimentos em mídias sociais.** 2015. 29 f. - Curso de Sociedade Brasileira de Computação, Sociedade Brasileira de Computação, Universidade Federal de Minas Gerais - MG, 2015.

BOYD, Danah M.; ELLISON, Nicole B. **Social network sites: Definition, history, and scholarship.** Journal of computer-mediated Communication, v. 13, n. 1, p. 210-230, 2007.

BRUCE, Rebecca. **A Bayesian Approach to Semi-Supervised Learning.** 2001. 8 f. Tese (Doutorado) - Computer Science, University Of North Carolina, Asheville, 2001. Disponível em: <http://www.afnlp.org/archives/nlprs2001/pdf/0100-01.pdf>. Acesso em: 22 maio 2022.

CARVALHO, Flavio et al. **Evaluating the Brazilian Portuguese version of the 2015 LIWC Lexicon with sentiment analysis in social networks.** 2019. 11 f. Tese (Doutorado) - Curso de Ciência de Dados, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2019.

CARVALHO, Paula; SILVA, Mário J. **SentiLex-PT: Principais características e potencialidades.** Oslo Studies in Language, v. 7, n. 1, 2015.

COLATRELLA, Carol. **Gender Codes: Why Women Are Leaving Computing** by Thomas J. Misa. Technology And Culture. Baltimore, p. 1-2. jan. 2012.

FONTANA, Eliton. **Introdução aos algoritmos de aprendizagem supervisionada.** Universidade Federal do Paraná: Departamento de Engenharia Química, 2020. Disponível em: [https://fontana.paginas.ufsc.br/files/2018/03/apostila\\_ML\\_pt2.pdf](https://fontana.paginas.ufsc.br/files/2018/03/apostila_ML_pt2.pdf) Acesso em: 11 jan de 2022.

GARDNER, Sue. **Why women are leaving the tech industry in droves.** Los Angeles Times. Los Angeles, p. 1-4. fev. 2014.

GIMPEL, Kevin; SCHNEIDER, Nathan; O'CONNOR, Brendan; DAS, Dipanjan; MILLS, Daniel; EISENSTEIN, Jacob; HEILMAN, Michael; YOGATAMA, Dani; FLANIGAN, Jeffrey; SMITH, Noah A.. **Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments.** 2010. 6 f. - School of Computer Science, Carnegie Mellon University, Pittsburgh, 2010.

GO, Alec; BHAYANI, Richa; HUANG, Lei. **Twitter Sentiment Classification using Distant Supervision.** 2009. 6 f. Stanford University, CA, 2009.

HEWLETT, Sylvia Ann; SERVON, Lisa J.; SHERBIN, Laura; SHILLER, Peggy; SOSNOVICH, Eytan; SOSNOVICH, Eytan. **The Athena Factor: Reversing the Brain Drain in Science, Engineering, and Technology.** Harvard Business Review. Boston, p. 1-109. jan. 2008.

HU, Minqing; LIU, Bing. **Mining and summarizing customer reviews.** Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 168-177, 2004

HUSSEIN, Doaa Mohey El-din Mohamed. **A survey on sentiment analysis challenges.** Journal Of King Saud University: Engineering Sciences. Cairo, Egito, 9 p. abr. 2016

HUTTO, Cj; GILBERT, Eric. **VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text.** 2014. 10 f. Tese (Doutorado) - Computer Science, Technology, Georgia Institute Of Technology, Atlanta, 2014.

IMTIAZ, Nasif; MIDDLETON, Justin; CHAKRABORTY, Joymallya; ROBSON, Neill; BAI, Gina, MURPHY-HILL, Emerson. **Investigating the effects of gender bias on GitHub.** In:

2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE). IEEE, 2019. p. 700-711.

JAVA, Akshay; SONG, Xiaodan; FININ, Tim; TSENG, Belle. Why we twitter: understanding microblogging usage and communities. In: ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 13., 2007, San Jose, California. **Why we twitter: understanding microblogging usage and communities**. New York: ACM, 2007. p. 56-65

JURAFSKY, Daniel; MARTIN, James H. **Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition**. Englewood Cliffs, New Jersey: Prentice Hall, 2000. 934 p.

LIU, Bing. **Sentiment Analysis**. New York, USA: Cambridge University Press, 2015. 769 p.

LIU, Bing. **Sentiment Analysis and Opinion Mining**. [s.l.]: Morgan & Claypool Publishers, 2012. 168 p.

LIU, Bing. **Sentiment Analysis and Subjectivity**. In: INDURKHAYA, Nitin; DAMERAU, Fred J. (Ed.). Handbook of Natural Language Processing. 2. ed. Boca Raton, Fl: Chapman and Hall/crc, 2010. Cap. 26. p. 627-666.

LIWC (org.). **Dictionary**. 2022. Disponível em: <https://www.liwc.app/dictionaries>. Acesso em: 16 maio 2022.

MELO, Randerson Lessa. **AVALIANDO O DICIONÁRIO EM PORTUGUÊS DO MÉTODO DE ANÁLISE DE SENTIMENTOS SENTISTRENGTH**. 2017. 55 f. TCC (Graduação) - Curso de Engenharia de Software, Universidade Federal do Ceará, Quixada, 2017.

NASKAR, Debashis; SINGH, Sanasam Ranbir; KUMAR, Durgesh; NANDI, Sukumar; RIVAHERRERA, Eva Onaindia de la. **Emotion Dynamics of Public Opinions on Twitter**. 2020. 24 f. Tese (Doutorado) - Acm Transactions On Information Systems (Tois),2013. p. 86-96

NLTK (org.). **Natural Language Toolkit**. 2022. Disponível em: <https://www.nltk.org/>. Acesso em: 16 maio 2022.

NORVIG, Peter. **How to Write a Spelling Corrector**. 2016. Disponível em: <<http://norvig.com/spell-correct.html>>. Acesso em: 26 out. 2022.

POSTGRESQL GLOBAL DEVELOPMENT GROUP (org.). **About PostgreSQL**. 2022. Disponível em: <https://www.postgresql.org/>. Acesso em: 16 maio 2022.

PYTHON (org.). **About Python**. 2022. Disponível em: <https://www.python.org/about/>. Acesso em: 16 maio 2022.

ORTU, Marco; DESTEFANIS, Giuseppe; COUNSELL, Steve; SWIFT, Stephen; TONELLI, Roberto; MARCHESI, Michele. **How diverse is your team? Investigating gender and nationality diversity in GitHub teams**. 2017. 18 f. - Journal Of Software Engineering Research And Development, v. 5, n. 1, p. 1-18, 2017.

OWOPUTI, Olutobi; O'CONNOR, Brendan; DYER, Chris; GIMPEL, Kevin; SCHNEIDER, Nathan; SMITH, Noah A.. **Improved Part-of-Speech Tagging for Online Conversational Text with Word Clusters**. 2013. 11 f. - School Of Computer Science, Carnegie Mellon University, Pittsburgh, 2013.

SAIF, Hassan; FERNANDEZ, Miriam; HE, Yulan; ALANI, Harith. **On Stopwords, Filtering and Data Sparsity for Sentiment Analysis of Twitter**. 2012. 8 f. - School Of Engineering And Applied Science, Aston University, UK, 2012.

SAIF, Hassan; FERNANDEZ, Miriam; HE, Yulan; ALANI, Harith. **Alleviating Data Sparsity for Twitter Sentiment Analysis**. 2012. 8 f. - Knowledge Media Institute, The Open University, UK, 2012.

SILVA, Nadia Felix Felipe da. **Análise de sentimentos em textos curtos provenientes de redes sociais**. 2016. 138 f. - Curso de Ciências de Computação e Matemática Computacional, São Paulo, 2016.

SPACY (org.). **SpaCy**. 2022. Disponível em: <https://spacy.io/>. Acesso em: 16 novembro 2022.

TABOADA, Maite; BROOKE, Julian; TOFILOSKI, Milan; VOLL, Kimberly; STEDE, Manfred. **Lexicon-Based Methods for Sentiment Analysis**. 2011. 42 f. - Computer Science, Department Of Linguistics, University Of Toronto, Toronto, 2011.

TAUSCZIK, Yla R.; PENNEBAKER, James W.. **The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods**. 2010. 32 f. Journal Of Language And Social Psychology, v. 29, n. 1, p. 24-54, 2010.

TWITTER (org.). **Twitter API**. 2022. Disponível em: <https://developer.twitter.com/en/products/twitter-api>. Acesso em: 16 maio 2022.

WORDNET. **About WordNet**. 2022. Disponível em: <https://wordnet.princeton.edu/>. Acesso em: 16 maio 2022.

JIANXING, Yu; ZHENG-JUN Zha; MENG, Wang, TAT-SENG, Chua. **Aspect ranking: Identifying important product aspects from online consumer reviews**. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1.1496-1505, 2011,

ZHENG, Lei; HAN, Kai. **Extracting Categorical Topics from Tweets Using Topic Model**. 2013. 11 f. N, Asia Information Retrieval Symposium, N, Springer, Berlin, Heidelberg, 2013. p. 86-96.



APÊNDICE B - Tabela de todos os aspectos extraídos e sua respectiva frequência.

<b>Aspecto</b>	<b>Frequência</b>
gente	569
pessoas	278
peessoa	125
coisa	93
cara	91
mundo	79
galera	69
dia	62
peessoal	61
empresa	60
problema	58
coisas	52
isso	50
vida	47
amiga	46
esse	42
tipo	39
mulher	39
empresas	38
deus	38
mãe	38
amigo	35

parte	34
maioria	32
evento	31
povo	31
trabalho	31
esses	31
curso	30
pai	28
tempo	28
ideia	28
inscrições	27
negócio	28
caras	25
homem	24
mulheres	24
uso	23
projeto	22
sonho	22
bolha	22
mano	22
você	21
ponto	21
essas	21
km	20
processo	20

lugar	19
dica	19
nome	18
parabéns	16
comunidade	15
verdade	15
javascript	15
país	15
sopa	15
twitter	15
mercado	15
falta	14
filha	14
essa	14
live	14
mão	14
jogo	13
noivo	13
objetivo	13
resto	12
pergunta	12
semana	12
livro	12
vídeo	12
fato	12

assédio	12
anos	11
front	11
programação	11
culpa	11
momento	11
resposta	11
bicho	11
código	11
área	11
time	11
amor	11
dono	11
link	11
amigos	11
vaga	11
rede	11
questão	11
algoritmo	10
então	10
pior	10
rolo	10
homens	10
banco	10
mina	10

honra	10
bom	10
salário	10
realidade	10
show	10
fio	10
suficiente	10
saudades	9
classe	9
pena	9
vamo	9
conteúdo	9
programa	9
meninas	9
dias	9
coração	9
regra	9
base	9
one	9
opção	9
bolo	9
caso	9
cidade	9
passo	8
luta	8

palavra	8
linguagem	8
chance	8
site	8
tipos	8
conceito	8
presidente	8
teste	8
solução	8
hora	8
quantidade	8
dinheiro	8
post	8
atenção	8
país	8
menina	8
tecnologia	8
merda	8
modelo	8
dicas	8
experiência	8
palestra	8
assunto	8
encontro	8
ciclo	8

minas	8
rapaz	8
perfil	8
projetos	8
palavras	8
episódio	8
fase	8
vezes	8
caminho	8
cardio	7
kit	7
papo	7
estudo	7
bora	7
brincadeira	7
café	7
luz	7
jornada	7
bateria	7
colega	7
lado	7
forma	7
raça	7
ser	7
qualidade	7

corpo	7
comida	7
irmã	7
conselho	7
funcionários	7
clã	7
governo	7
cabelo	7
turma	7
história	7
lista	7
lúcida	7
aulas	7
email	7
eventos	7
desculpa	7
equipe	7
juro	7
artigo	7
tweet	7
esquenta	7
alias	7
musa	6
dona	6
testes	6

opinião	6
ego	6
versão	6
marca	6
linha	6
sol	6
amigas	6
diferença	6
exemplo	6
argumento	6
entrevista	6
redes	6
reunião	6
importante	6
sênior	6
produto	6
computador	6
celular	6
firma	6
época	6
gatos	6
construção	6
matrículas	6
tanto	6
programo	6

transição	6
meta	6
painel	6
abraço	6
tia	6
médico	6
medo	6
vontade	6
série	6
moça	6
criança	6
família	6
cor	6
ideal	6
voto	6
liderança	6
remuneração	6
dinâmica	6
feedback	6
mentiras	6
motivo	5
comentário	5
metade	5
dúvida	5
segredo	5

leitura	5
intenção	5
chefe	5
ferramentas	5
professora	5
professores	5
mel	5
problemas	5
real	5
spoiler	5
gringo	5
explicação	5
cliente	5
ódio	5
saudade	5
diabo	5
restaurante	5
bruxa	5
mail	5
nosa	5
bola	5
senhora	5
fã	5
sistemas	5
gosto	5

dor	5
marido	5
básico	5
gestão	5
casa	5
oportunidade	5
discurso	5
maluco	5
aplicativo	5
mentira	5
iniciativas	5
rainha	5
ano	5
cenário	5
filme	5
módulo	5
juntas	5
conversa	5
mês	5
mentora	5
plataforma	5
troca	5
sorteio	5
foto	5
cursos	5

vídeos	5
ferramenta	5
trem	5
golpe	5
preço	5
preocupação	5
aros	5
cabeça	5
definição	5
comigo	5
treino	5
bolsonaro	5
ponte	5
situação	5
vitória	5
bug	5
frase	5
benefícios	5
responsabilidade	5
idade	5
super	5
resultado	5
gestor	5
aplicação	5
verme	5

pele	4
cargo	4
contato	4
motivação	4
pop	4
professor	4
inferno	4
recrutadores	4
fonte	4
verdades	4
campo	4
vó	4
abstração	4
contexto	4
cachorro	4
necessidade	4
servidor	4
meses	4
shape	4
performance	4
mensagem	4
transmissão	4
layout	4
avó	4
ovo	4

remédio	4
detalhe	4
curiosidade	4
informação	4
futuro	4
cabos	4
câmera	4
investimento	4
água	4
motorista	4
usuário	4
cérebro	4
música	4
organização	4
academia	4
relacionamento	4
flores	4
proposta	4
jovens	4
pandemia	4
tema	4
estratégia	4
conto	4
engenheira	4
passo	4

inscrição	4
ingressos	4
vagas	4
lista	4
bots	4
grupo	4
democracia	4
teoria	4
teclado	4
chances	4
mapa	4
informações	4
grade	4
login	4
resumo	4
plano	4
dados	4
porta	4
comentários	4
pesquisa	4
roupas	4
melhora	4
dívida	4
tratamento	4
pincel	4

aliens	4
partes	4
escolha	4
linkedin	4
educação	4
clima	4
regulamentação	4
área	4
desafios	4
estágio	4
decisão	4
carreira	4
população	4
leiteiro	4
notícia	4
voz	4
silêncio	4
analista	4
experiências	4
resultados	4
moço	4
macho	4
sorte	4
top	4
comportamento	4

podcast	4
prefeito	4
treinamento	4
caos	3
mentiria	3
afirmação	3
alegria	3
via	3
documentação	3
monstro	3
musica	3
gangs	3
monte	3
black	3
conteudo	3
salários	3
atendente	3
recrutadora	3
scan	3
geração	3
gato	3
prédio	3
valor	3
github	3
seguidores	3

número	3
vizinho	3
maravilha	3
estado	3
camisa	3
obrigada	3
frutas	3
lance	3
bloco	3
fim	3
programador	3
desenvolvedor	3
desafio	3
feira	3
bolinho	3
agenda	3
poder	3
profissionais	3
fatores	3
namorado	3
foco	3
online	3
taxa	3
namorada	3
satisfação	3

questionamento	3
tristeza	3
fãs	3
noite	3
tinto	3
céu	3
áreas	3
drama	3
amores	3
gole	3
brasileiros	3
humanidade	3
estrutura	3
entrada	3
torres	3
edição	3
sentimento	3
impacto	3
sucesso	3
valores	3
temporada	3
questões	3
conhecimento	3
receita	3
pá	3

deusa	3
aniversário	3
emprego	3
injeção	3
caminhão	3
clássico	3
sensação	3
primo	3
alunas	3
técnica	3
estação	3
documento	3
aplicações	3
pé	3
graus	3
autor	3
arquivo	3
entregador	3
rota	3
blog	3
postura	3
currículo	3
ansiedade	3
mana	3
impressão	3

existência	3
pão	3
objeto	3
espaço	3
debate	3
interesse	3
conexão	3
camiseta	3
padre	3
engajamento	3
códigos	3
caminhos	3
codigo	3
tarefa	3
noticias	3
ato	3
alunos	3
melhoras	3
abordagem	3
menino	3
família	3
público	3
ambiente	3
maneira	3
formas	3

destino	3
rua	3
vacinas	3
obrigação	3
banda	3
lenda	3
data	3
eleições	3
gerente	3
tech	3
loucura	3
caro	3
perdão	3
delícia	3
ar	3
ordem	3
férias	3
missão	3
conteúdos	3
energia	3
garçom	3
treta	3
companheiro	3
termo	3
referências	3

parcela	3
texto	3
idades	3
bolsonarismo	3
máscara	3
roupa	3
neto	3
pesquisas	3
galeria	3
procura	3
posição	3
candidato	3
socorro	3
perguntas	3
síndrome	3
eleitores	3
grupos	3
vivo	3
workshop	3
pilar	3
nômades	3
times	3
parada	3
obra	3
orgulho	3

ausência	3
coroa	3
fator	3
prima	2
importância	2
vez	2
carteira	2
tarefas	2
páginas	2
funcho	2
parente	2
nível	2
salario	2
facebook	2
apartamento	2
jornalismo	2
hacker	2
padrão	2
bolhadev	2
principio	2
brinde	2
coreto	2
tentativa	2
maneiro	2
fratura	2

referência	2
cena	2
uva	2
média	2
escopo	2
foguete	2
maias	2
amizades	2
exercício	2
pipoca	2
opções	2
sites	2
pipeline	2
inspiração	2
recorde	2
lobo	2
atualização	2
carga	2
colo	2
situações	2
placa	2
estabilidade	2
ninho	2
produtividade	2
desconto	2

final	2
pros	2
candidatos	2
domingo	2
faculdade	2
papai	2
tamanho	2
copia	2
clientes	2
lives	2
mínimo	2
estagiário	2
peso	2
padaria	2
refeições	2
atleta	2
sexta	2
barulho	2
volta	2
torcida	2
discussão	2
caneta	2
lâmpada	2
plug	2
hub	2

iniciante	2
contas	2
preconceito	2
elenco	2
reforma	2
conclusão	2
desejo	2
camarada	2
bar	2
adolescente	2
gaze	2
festa	2
sentido	2
loja	2
escritório	2
universitários	2
bagulho	2
participantes	2
obstáculos	2
crise	2
expectativa	2
louça	2
segundos	2
sonoridade	2
vírus	2

colegas	2
exército	2
windows	2
note	2
fome	2
chave	2
solo	2
inteligência	2
máquina	2
timing	2
patrocinadores	2
cadeiras	2
convite	2
feriado	2
gratidão	2
inovação	2
formato	2
deusas	2
semestre	2
tour	2
gentes	2
chocolate	2
contrato	2
jornalista	2
ônibus	2

ilusão	2
controle	2
cadeira	2
mar	2
cansaço	2
diretora	2
sangue	2
dificuldades	2
entrevistas	2
discussões	2
lançamento	2
relação	2
mudanças	2
buraco	2
comando	2
saúde	2
minorias	2
artigos	2
imposto	2
custo	2
caixeiro	2
diferenças	2
obras	2
arquivos	2
prática	2

médica	2
ouro	2
tatuagem	2
mães	2
sotaque	2
lugares	2
motivos	2
painéis	2
frustração	2
praia	2
altura	2
rancor	2
paciência	2
anjo	2
tiro	2
bunda	2
investidores	2
voltas	2
chato	2
perfis	2
personagens	2
fenômeno	2
capitalismo	2
signo	2
efeito	2

respeito	2
esquerda	2
zona	2
sumo	2
atualizações	2
descrição	2
papel	2
fogo	2
pedido	2
nordestino	2
estudantes	2
design	2
método	2
ferias	2
ressaca	2
idiomas	2
bolero	2
vacina	2
entrega	2
revisão	2
fruta	2
capacidade	2
sul	2
cd	2
criadores	2

justificativa	2
dificuldade	2
sinceridade	2
vigilante	2
fotos	2
exame	2
jornalistas	2
ciência	2
oposto	2
pato	2
demanda	2
sistema	2
gasolina	2
espaços	2
transportadora	2
manager	2
habilidades	2
update	2
mídia	2
veia	2
jogador	2
copa	2
aula	2
policial	2
dependência	2

endereço	2
presença	2
relatório	2
pesquisadores	2
protesto	2
raço	2
nomes	2
romana	2
horas	2
casal	2
emoção	2
capa	2
travesti	2
ecofascismo	2
pesadelo	2
histórias	2
autora	2
amizade	2
obsessão	2
óculos	2
gays	2
execução	2
grupinho	2
emoji	2
beleza	2

treinamentos	2
consistência	2
inseguranças	2
reflexões	2
moda	2
incêndio	2
função	2
sorvete	2
monitor	2
profissional	2
metabolismo	2
vítima	2
inglês	2
percepção	2
aquisição	2
diabos	2
atração	2
relações	2
moral	2
senhoras	2
vantagem	2
pronúncia	2
cartão	2
rádio	2
retorno	2

dieta	2
psicólogo	2
terapia	2
novidade	2
peça	2
senhor	2
pílula	2
corda	2
ingresso	2
visto	2
software	2
apoio	2
brasileiro	2
seres	2
morada	2
cheiro	2
estados	2
som	2
startup	2
dólar	2
vaso	2
usuários	2
instância	2
famílias	2
casos	2

dilema	2
frio	2
polícia	2
fantasia	2
filhos	2
cachorra	2
liberdade	2
irmão	2
cachorros	2
flag	2
quadro	2
criança	2
diretor	2
realização	2
casas	2
falas	2
tio	2
bizarro	2
lesões	2
exemplos	2
barba	2
saco	2
recrutador	2
praga	2
lembrete	2

bonito	2
mentalidade	2
pontos	2
inata	2
gestores	2
mediadores	2
bichinho	2
rim	2
sábado	2
rosto	2
contratação	2
líder	2
momentos	2
fraude	2
policia	2
video	2
respostas	2
minions	2
expert	2
agosto	2
janelas	2
workshops	2
embaixadora	2
práticas	2
dúvidas	2

trilha	2
palestras	2
amazona	2
droga	2
imagens	2
isolamento	2
efeitos	2
mapeamento	2
covil	2
mito	2
escola	2
contrário	2
distância	2
eclipse	2
idea	2
incompetência	2
legal	2
audio	2
lições	2
coluna	2
vazamento	2
servidores	2
métodos	2
auge	2
mala	2

marketing	1
álbum	1
notificações	1
transação	1
inconsistência	1
vidro	1
fiscais	1
moderador	1
calcinha	1
bandidos	1
movimento	1
typescript	1
emails	1
bolhas	1
transparência	1
poster	1
engenheiro	1
leituras	1
gringos	1
melhor	1
lança	1
upgrade	1
compilado	1
revolucionário	1
livros	1

favorito	1
criador	1
porcentagem	1
circuitos	1
hardware	1
rastilho	1
navegadores	1
controladora	1
cambiar	1
militante	1
trigonometria	1
papagaios	1
psicologo	1
estreante	1
games	1
comunidades	1
biblioteca	1
picha	1
massagem	1
bosta	1
utilidade	1
conversão	1
bobagem	1
engenharia	1
coesão	1

marceneiro	1
exploração	1
fontes	1
mudos	1
moedas	1
operadores	1
monstros	1
riqueza	1
operações	1
expressão	1
nerd	1
lenço	1
ataque	1
demora	1
baixa	1
audácia	1
castanha	1
avaliação	1
branco	1
gatinhos	1
papeis	1
áudio	1
validação	1
inicial	1
pendrive	1

cordas	1
fama	1
mostras	1
spotify	1
quebra	1
clube	1
robô	1
salgadinhos	1
churrasco	1
campeão	1
beca	1
ânimo	1
esgoto	1
atividade	1
ave	1
farinha	1
distração	1
skin	1
babaca	1
nave	1
maconha	1
humilhação	1
pane	1
guerreiros	1
misericórdia	1

astro	1
especificação	1
mictório	1
povinho	1
boatos	1
pretensão	1
tabagismo	1
bot	1
hashtag	1
busto	1
registro	1
sofá	1
astros	1
insônia	1
bonde	1
aleatoriedade	1
reação	1
nervosismo	1
abstinência	1
adesivo	1
maldições	1
sensibilização	1
vizinhos	1
união	1
mato	1

lençol	1
pregadores	1
fumante	1
herdeiros	1
mesinha	1
gerenciamento	1
gravei	1
tampa	1
tábua	1
ganho	1
pubs	1
glóbulos	1
jedi	1
larica	1
cuscutz	1
revelação	1
bônus	1
pair	1
tempos	1
lição	1
canal	1
bagel	1
freguês	1
gafes	1
bonequinha	1

certificado	1
burocracia	1
aroma	1
apresentador	1
minutos	1
incas	1
proteína	1
especie	1
fobia	1
precioso	1
seniores	1
cadastro	1
marombeira	1
curiosidades	1
esquema	1
lojas	1
camarote	1
equilíbrio	1
mimeógrafo	1
combo	1
nego	1
vovô	1
petição	1
farol	1
posto	1

portas	1
aspirador	1
apresentação	1
ti	1
reis	1
bagunça	1
generalista	1
componentes	1
confiabilidade	1
serviço	1
prateleira	1
flexibilização	1
apresentadora	1
barbada	1
almoço	1
coragem	1
cloud	1
micro	1
acordem	1
barramento	1
escrotos	1
carroça	1
símbolo	1
parafusos	1
tintas	1

editores	1
prioridade	1
juntos	1
morte	1
saladeira	1
softwares	1
imersão	1
brotheragem	1
softskill	1
castração	1
rapidez	1
pontas	1
contras	1
cachos	1
sorteios	1
potênciastech	1
análise	1
parceira	1
consultoria	1
celebração	1
indicação	1
amantes	1
cientistas	1
legado	1
matemáticos	1

relevância	1
youtube	1
passador	1
repertório	1
trajetória	1
tirado	1
universo	1
confiança	1
começo	1
chuveiro	1
miopia	1
músculo	1
teia	1
hospital	1
hobby	1
laranja	1
felicidades	1
votação	1
campeã	1
elegância	1
novembro	1
mecanismo	1
resolução	1
língua	1
cereja	1

reposição	1
formação	1
sabor	1
raiva	1
desbravadores	1
personificação	1
bordel	1
obstáculo	1
empresto	1
campeões	1
glória	1
poste	1
alma	1
depoimentos	1
anemia	1
multidão	1
metas	1
macroregião	1
organizações	1
coletiva	1
operador	1
credenciais	1
repositório	1
primavera	1
jaqueta	1

cursor	1
manifesto	1
comandos	1
raiz	1
mercadoria	1
gravações	1
engraçado	1
desktop	1
consumidora	1
estruturas	1
marco	1
pagamento	1
lembrança	1
tanque	1
águas	1
magia	1
migrações	1
tabelas	1
possível	1
psicóloga	1
funcionária	1
prematura	1
modernização	1
interpretações	1
danos	1

tempestade	1
algoritmos	1
museu	1
registros	1
turismo	1
reset	1
filantrópo	1
exploração	1
microserviços	1
buscas	1
árvores	1
aurora	1
piadas	1
comediante	1
logo	1
colocação	1
séculos	1
pasta	1
paternidade	1
certificação	1
visualizações	1
impressões	1
cidadania	1
arrependimento	1
mimos	1

palpite	1
ingredientes	1
cisne	1
hipoteca	1
ranking	1
escovas	1
secador	1
eletrodomésticos	1
dose	1
milagre	1
framework	1
selfie	1
noiva	1
boxe	1
clone	1
capricho	1
materiais	1
sinais	1
observador	1
sinusite	1
direito	1
complexo	1
eleição	1
importa	1
eleitor	1

massacre	1
americanos	1
templo	1
canção	1
nojo	1
continentes	1
historia	1
aviadores	1
original	1
camareiro	1
risada	1
elo	1
bilionários	1
galinha	1
indagação	1
descendente	1
admiração	1
drogas	1
doença	1
cheques	1
matraca	1
metaverso	1
cripto	1
raptos	1
acesso	1

capitalista	1
pontuação	1
defeito	1
bloqueio	1
gentalha	1
inimigos	1
marcas	1
crítica	1
finalidade	1
arranjo	1
mula	1
conta	1
habilidade	1
normal	1
personal	1
documentário	1
cassino	1
completo	1
fascista	1
pingo	1
namoro	1
músicas	1
caravana	1
exatas	1
genes	1

credibilidade	1
humanos	1
eternidade	1
frutos	1
fascismo	1
melatonina	1
desrespeito	1
disputa	1
piada	1
gêmeas	1
reboco	1
boletim	1
religiões	1
análises	1
coelho	1
ladainha	1
compromisso	1
carapuça	1
iphone	1
boticário	1
aplicativos	1
cílios	1
propriedade	1
objetiva	1
expectativas	1

inimigo	1
macarrão	1
escrito	1
matemática	1
chapéuzinho	1
prazos	1
técnicas	1
raiozinho	1
zoo	1
distinção	1
promoções	1
licor	1
preferências	1
mensagens	1
áudios	1
iniciantes	1
precisa	1
cenas	1
van	1
bolos	1
consequências	1
estilo	1
solteira	1
alienígenas	1
sugestão	1

consciência	1
hélio	1
vergonha	1
bandeira	1
saudação	1
fraturas	1
remédios	1
farmácia	1
fantasias	1
vento	1
task	1
comidas	1
cerimônia	1
busca	1
gastrite	1
garota	1
propaganda	1
marcos	1
gelatina	1
único	1
manutenção	1
terminologia	1
foices	1
campus	1
flexão	1

analistas	1
crescimento	1
imagem	1
acidente	1
roubo	1
mesada	1
funcionalidade	1
linguagens	1
pensamento	1
prisão	1
cuidado	1
superstar	1
abreviação	1
roleta	1
ratinho	1
bruxas	1
justiça	1
merdas	1
fada	1
cereal	1
obvio	1
analisador	1
propriedades	1
professoras	1
letras	1

complexidade	1
trato	1
fofoca	1
excesso	1
nível	1
musico	1
medico	1
camadas	1
trens	1
drácula	1
mensalidade	1
vestibular	1
telefones	1
diferenciação	1
condição	1
legumes	1
cones	1
propósito	1
licença	1
perolas	1
trabalhador	1
provas	1
responsabilidades	1
grafismo	1
métrica	1

parceiro	1
origem	1
conflitos	1
descentralização	1
ritmo	1
certeza	1
observações	1
narinas	1
dicção	1
maquinista	1
jeito	1
inteligencia	1
seleção	1
substituições	1
assassinos	1
transplante	1
advogada	1
moleque	1
disparidades	1
categorias	1
termos	1
relato	1
escolas	1
abata	1
brasileira	1

encerramento	1
historiadora	1
picolé	1
atitudes	1
fervo	1
povos	1
canto	1
chefes	1
financiamento	1
iniciativa	1
quadrinho	1
quarto	1
guias	1
orientadora	1
picaretas	1
maré	1
próxima	1
pretos	1
urnas	1
retórica	1
nazismo	1
laboratório	1
vingadores	1
mudança	1
racistas	1

funcionário	1
operação	1
queda	1
fronha	1
karaokê	1
poderosos	1
sensacional	1
guerreiro	1
lei	1
graça	1
major	1
coletividade	1
pastor	1
contribuições	1
envolvimento	1
ambicioso	1
negros	1
natal	1
confusões	1
prostituição	1
general	1
desinformação	1
console	1
estoque	1
filmes	1

crepioca	1
conhecido	1
ajuda	1
sintoma	1
estereótipo	1
topo	1
preguiça	1
retenção	1
alternador	1
mecânico	1
tecla	1
rotina	1
pãozinho	1
mingau	1
sulistas	1
considerações	1
panorama	1
espiã	1
sons	1
joelhada	1
religião	1
tela	1
beijo	1
acaso	1
surto	1

segurança	1
garganta	1
entretenimento	1
suportes	1
dominação	1
condições	1
frequência	1
frente	1
americano	1
bombeiro	1
prisma	1
seringa	1
título	1
provocação	1
indiana	1
casamenteira	1
máfia	1
gafe	1
gabinete	1
links	1
migração	1
futuros	1
levantamento	1
cachaça	1
tampo	1

fungos	1
madeira	1
cozinha	1
nova	1
conjunto	1
unha	1
mesas	1
engenheiros	1
tampos	1
merengues	1
kilos	1
risco	1
solidão	1
moças	1
bandido	1
policimento	1
essência	1
transformação	1
sindicato	1
boato	1
cativeiro	1
abdômen	1
ciúmes	1
desespero	1
bandas	1

bullying	1
divórcio	1
quadrilha	1
panturrilhas	1
pedreiro	1
feiras	1
bíceps	1
componente	1
click	1
apagador	1
ligação	1
barreiras	1
senha	1
renda	1
vertebra	1
olho	1
cozinheiro	1
oportunidades	1
diferencial	1
volume	1
negligência	1
experiecia	1
negativa	1
espécie	1
frango	1

bluetooth	1
transferência	1
brownie	1
pedreiros	1
saldo	1
conselhos	1
telas	1
esteroides	1
risca	1
esteróides	1
indústria	1
formulários	1
bibliotecas	1
natureza	1
acordo	1
casamento	1
individualidade	1
derivados	1
gostinho	1
influência	1
flacidez	1
musculatura	1
circulação	1
calendários	1
veteranos	1

soldado	1
semanas	1
tenis	1
mané	1
porcaria	1
minutinhos	1
floridas	1
venda	1
chão	1
espinha	1
preços	1
árvore	1
hemisfério	1
alimentos	1
discursos	1
gol	1
sacos	1
armas	1
vizinhança	1
evolução	1
secretaria	1
fulo	1
prova	1
observação	1
efêmero	1

agregado	1
limite	1
cais	1
drive	1
tecnica	1
autônomo	1
juiza	1
ladrão	1
garotos	1
universidade	1
provedores	1
calças	1
pássaro	1
lentidão	1
especialistas	1
transportes	1
saleiro	1
visão	1
diretoria	1
eleitorado	1
caráter	1
agência	1
censo	1
administração	1
instâncias	1

título	1
capitão	1
ideologia	1
pátria	1
orçamento	1
etapas	1
constituição	1
luto	1
supermercados	1
imunidade	1
capítulo	1
vulnerabilidade	1
plenário	1
lutador	1
caminhoneiros	1
bones	1
nordeste	1
denúncias	1
campanha	1
atriz	1
tabela	1
espécime	1
urina	1
machado	1
genocídio	1

cura	1
pronomes	1
personalidade	1
causa	1
violinistas	1
forças	1
cão	1
descontentamento	1
empreendimento	1
afirmações	1
aconchego	1
indícios	1
buffet	1
mangueira	1
consumidor	1
produtos	1
brisa	1
esquilos	1
sócia	1
vitamina	1
aparição	1
absurdo	1
desfile	1
senso	1
elevador	1

diária	1
chá	1
idéia	1
imbecil	1
etimologia	1
cidadao	1
cantor	1
dinâmicas	1
taça	1
estomago	1
desigualdade	1
peles	1
tumor	1
piercing	1
cartilagem	1
amarelo	1
carne	1
babado	1
narrativa	1
declaração	1
discoteca	1
preparo	1
mordida	1
hormônios	1
bases	1

finanças	1
oculos	1
bolso	1
símbolos	1
misterio	1
ameixa	1
cenários	1
sindical	1
finalização	1
parágrafos	1
famigerado	1
alienação	1
patriotas	1
adversário	1
padrasto	1
violência	1
etapa	1
restruturação	1
produções	1
gracinha	1
asma	1
atual	1
respiro	1
gado	1
crente	1

baixo	1
presidência	1
documentos	1
premium	1
repercussão	1
indivíduos	1
escada	1
juntinhos	1
bastidores	1
inteira	1
estudante	1
lavagem	1
programas	1
esperado	1
gatilho	1
peças	1
recomendações	1
conexões	1
início	1
abril	1
concorrência	1
inveja	1
histórico	1
personalidades	1
calor	1

profissão	1
dores	1
horário	1
exposto	1
atribuições	1
dev	1
procrastinação	1
coruja	1
operário	1
denúncia	1
agentes	1
policiais	1
prefeitura	1
ditadura	1
hétero	1
notebook	1
guria	1
exceções	1
carças	1
infraestrutura	1
detalhes	1
empregos	1
mercado	1
militância	1
corrupção	1

mentor	1
estudos	1
rocha	1
advogadas	1
estagiária	1
pauta	1
prêmio	1
princesa	1
reino	1
limites	1
feixe	1
doação	1
face	1
relacionamentos	1
onda	1
escória	1
traumas	1
rodapé	1
depressão	1
tentativas	1
tigela	1
acesa	1
vendas	1
reta	1
empendedorismofeminino	1

identidades	1
física	1
voucher	1
lema	1
cargos	1
aprendizado	1
combate	1
estrutural	1
recrutamento	1
temas	1
reparação	1
futebol	1
gostasas	1
roda	1
potências	1
ação	1
atividades	1
salas	1
fundadoras	1
curadora	1
tensão	1
lastimo	1
multa	1
ápice	1
hábito	1

galho	1
lembro	1
jogos	1
cola	1
pacote	1
trio	1
afeto	1
personagem	1
salvador	1
submissões	1
paisagens	1
sono	1
receptividade	1
sorveteria	1
chapéu	1
querida	1
armação	1
algodão	1
braço	1
bobeira	1
chute	1
pirâmide	1
diálogo	1
esquecimento	1
memória	1

inverno	1
barato	1
velha	1
pagode	1
sintomas	1
banho	1
racismo	1
desigualdades	1
reportagem	1
block	1
cabelos	1
tortura	1
posicionamento	1
gerentes	1
restrição	1
médicos	1
economia	1
cônjuge	1
senhorio	1
porções	1
festas	1
filas	1
inflação	1
passagem	1
abelha	1

absurdos	1
palácio	1
costas	1
gota	1
vestido	1
botão	1
chuva	1
promessa	1
morangos	1
especialização	1
tecido	1
malevolência	1
suporte	1
manchete	1
saia	1
pureza	1
profissional	1
regras	1
prato	1
filhotes	1
bomba	1
trauma	1
companhias	1
elétrico	1
gasto	1

fabricante	1
taxi	1
explicações	1
países	1
carregadores	1
trajeto	1
estações	1
proibição	1
polêmica	1
relógio	1
monitoramento	1
cores	1
display	1
anel	1
entendimento	1
pescoço	1
percentual	1
oferta	1
nutricionista	1
cobertura	1
mágica	1
interpretação	1
pombo	1
apito	1
ensino	1

signos	1
diagnostico	1
articulações	1
flerte	1
parto	1
comissão	1
nublado	1
fotografia	1
guia	1
amendoim	1
necessidades	1
embuste	1
milho	1
catalogo	1
paciencia	1
reflexão	1
exercício	1
reserva	1
principal	1
indicações	1
força	1
comparativo	1
trecho	1
movimentação	1
apelo	1

variáveis	1
criação	1
server	1
problemática	1
vantagens	1
filosofia	1
página	1
politica	1
reatores	1
configurações	1
demonstração	1
artista	1
sapato	1
filho	1
ornitologia	1
generalização	1
soluções	1
alcance	1
contribuição	1
cigarro	1
adoção	1
público	1
tatuador	1
arquitetura	1
compra	1

inspirações	1
esperança	1
saga	1
tecnologias	1
plataformas	1
arroz	1
tópico	1
prototipo	1
discordo	1
combinado	1
suplementos	1
transporte	1
atributos	1
herança	1
relatividade	1
interface	1
hierarquia	1
refêns	1
perspectiva	1
provedor	1
acostagem	1
representação	1
companhia	1
assuntos	1
engano	1

novela	1
estabelecimento	1
rancho	1
desabafo	1
formatura	1
construtor	1
injetor	1

**Tabela 8:** Frequência dos aspectos coletados