

**UNIVERSIDADE DE CAXIAS DO SUL  
ÁREA DO CONHECIMENTO DE CIÊNCIAS EXATAS E  
ENGENHARIAS**

**ÂNDERSON ZORRER DALMINA**

**AVALIAÇÃO DO USO DE TRADUÇÃO DE MÁQUINA NA ANÁLISE  
DE SENTIMENTO MULTI-IDIOMAS**

**CAXIAS DO SUL**

**2022**

**ÂNDERSON ZORRER DALMINA**

**AVALIAÇÃO DO USO DE TRADUÇÃO DE MÁQUINA NA ANÁLISE  
DE SENTIMENTO MULTI-IDIOMAS**

Trabalho de Conclusão de Curso  
apresentado como requisito parcial à  
obtenção do título de Engenheiro de  
Computação na Área do Conhecimento  
de Ciências Exatas e Engenharias da  
Universidade de Caxias do Sul.

Orientador: Prof. Dr. André Gus-  
tavo Adami

**CAXIAS DO SUL**

**2022**

**ÂNDERSON ZORRER DALMINA**

**AVALIAÇÃO DO USO DE TRADUÇÃO DE MÁQUINA NA ANÁLISE  
DE SENTIMENTO MULTI-IDIOMAS**

Trabalho de Conclusão de Curso  
apresentado como requisito parcial à  
obtenção do título de Engenheiro de  
Computação na Área do Conhecimento  
de Ciências Exatas e Engenharias da  
Universidade de Caxias do Sul.

**Aprovado em 01/07/2022**

**BANCA EXAMINADORA**

---

Prof. Dr. André Gustavo Adami  
Universidade de Caxias do Sul - UCS

---

Prof. Dr. André Luis Martinotto  
Universidade de Caxias do Sul - UCS

---

Profa. Dra. Carine Geltrudes Webber  
Universidade de Caxias do Sul - UCS

## **AGRADECIMENTOS**

Agradeço a minha mãe, Elaine, e ao meu pai, Gilberto, por todo apoio, incentivo, amor e educação recebidos. À minha irmã, Natália, que compartilhou essa jornada comigo. Amo vocês. Essa conquista é de todos nós.

Agradeço a minha namorada, Paula, por todo amor, apoio e companheirismo durante todos esses anos.

Agradeço ao meu orientador, Prof. Dr. Andre Gustavo Adami, pelos ensinamentos e orientações durante todo meu processo de graduação e desenvolvimento desse trabalho.

Agradeço a todos os professores que contribuíram com minha jornada acadêmica pelos conhecimentos compartilhados.

Agradeço aos meus amigos e colegas, pelo companheirismo, tornando essa jornada mais leve.

Agradeço a todos que contribuíram na minha formação pessoal e profissional.

*“A maior parte dos sentimentos habita um terreno onde palavra alguma jamais pisou.”*

***Rainer Maria Rilke***

## RESUMO

Dado o aumento significativo da expressão de sentimentos em meios digitais, através de avaliações em plataformas online, a análise de sentimento de textos vem ganhando importância nas tarefas de processamento de linguagem natural. Porém, muitos dos métodos disponíveis para análise de sentimento são desenvolvidos para o idioma inglês. Diante dessa limitação, pode ser preciso utilizar diferentes abordagens para realizar a análise de sentimento em múltiplos idiomas. Uma dessas abordagens é a tradução de máquina. Dessa forma, esse trabalho visa avaliar diferentes abordagens de análise de sentimento para múltiplos idiomas utilizando o modelo BERT. Essa avaliação é realizada por meio do desenvolvimento de um sistema de análise de sentimento e execução de experimentos utilizando abordagem nativa e de tradução de máquina. A abordagem nativa tem por objetivo realizar a análise de sentimento utilizando métodos específicos para cada idioma trabalhado. A abordagem de tradução de máquina realiza a tradução de textos de diferentes idiomas para o inglês, e a análise de sentimento é realizada utilizando apenas métodos para o idioma inglês. Os idiomas analisados nesse trabalho são: inglês, alemão, espanhol e francês. Os resultados demonstraram que as abordagens nativas obtiveram melhor desempenho que a abordagem de tradução de máquina em 3 dos 4 idiomas analisados. Porém, dependendo dos requisitos, a diferença de resultado pode não ser significativa, podendo ser aplicável a abordagem de tradução de máquina.

**Palavras-chave:** Análise de sentimento. Multi-idiomas. Processamento de linguagem natural. Tradução. Aprendizado de máquina. *transformer*. BERT.

## LISTA DE FIGURAS

Figura 1 – Buscas por “ <i>Sentiment Analysis</i> ” no Google desde 2004 . . . . .	13
Figura 2 – Tarefas realizadas pela análise de sentimento . . . . .	14
Figura 3 – Etapas da análise de sentimento . . . . .	16
Figura 4 – Categorização das abordagens de análise de sentimento . . . . .	20
Figura 5 – Fases da abordagem baseada em léxico . . . . .	22
Figura 6 – Fases da abordagem de aprendizado de máquina supervisionado . . . . .	24
Figura 7 – Diagrama da abordagem híbrida . . . . .	24
Figura 8 – Arquitetura transformer . . . . .	26
Figura 9 – Mecanismo de atenção na arquitetura <i>transformer</i> . . . . .	27
Figura 10 – Processo de aplicação do vetor de posições . . . . .	28
Figura 11 – Etapas de pré-treinamento e <i>fine-tuning</i> do modelo BERT . . . . .	29
Figura 12 – Matriz de confusão para problemas de 2 classes . . . . .	31
Figura 13 – Abordagem multi-idioma baseada em léxico . . . . .	34
Figura 14 – Abordagem multi-idioma baseada em corpus paralelos . . . . .	35
Figura 15 – Abordagem multi-idioma baseada em aprendizado de máquina . . . . .	35
Figura 16 – Abordagem multi-idioma baseada em tradução de máquina . . . . .	36
Figura 17 – Valores de precisão, <i>recall</i> e F1-Score obtidos por método . . . . .	38
Figura 18 – Desempenho da análise de sentimento por tradução de máquina nos idiomas trabalhados. . . . .	42
Figura 19 – Quantidade de <i>tokens</i> gerados e sua distribuição . . . . .	45
Figura 20 – Diagrama da abordagem nativa . . . . .	47
Figura 21 – Diagrama da abordagem de tradução de máquina . . . . .	47
Figura 22 – Estrutura de cada avaliação na base de dados . . . . .	49

## LISTA DE TABELAS

Tabela 1 – Valores de acurácia e F1-Score obtidos por método. . . . .	39
Tabela 2 – Valores de acurácia e F1-Score obtidos por método e idioma. . . . .	40
Tabela 3 – Resultado da análise de sentimento em cada ferramentas de tradução de máquina. . . . .	41
Tabela 4 – Modelo <i>Bidirectional Encoder Representations from Transformers</i> (BERT) utilizado em cada idioma. . . . .	44
Tabela 5 – Estatísticas dos <i>tokens</i> gerados . . . . .	45
Tabela 6 – Comparação dos métodos de validação . . . . .	46
Tabela 7 – Estatísticas da base de treinamento. . . . .	49
Tabela 8 – Desempenho da abordagem nativa em cada idioma. . . . .	50
Tabela 9 – Desempenho da análise de sentimentos para as traduções de diferentes idiomas para o inglês. Os resultados para a língua inglesa são mostrados como referência. . . . .	51
Tabela 10 – Estatísticas dos <i>tokens</i> por avaliação das bases traduzidas para o inglês. . . . .	51
Tabela 11 – Desempenho das abordagens nativas e tradução de máquina . . . . .	52
Tabela 12 – Desempenho das abordagens quando avaliadas as polaridades separadamente. . . . .	53

## LISTA DE ABREVIATURAS E SIGLAS

<b>NLP</b>	<i>Natural Language Processing</i>
<b>API</b>	<i>Application Programming Interface</i>
<b>TF-IDF</b>	<i>Term Frequency, Inverse Document Frequency</i>
<b>CBOW</b>	<i>Continuous Bag of Words</i>
<b>SVM</b>	<i>Support Vector Machine</i>
<b>ANN</b>	<i>Artificial Neural Network</i>
<b>RNN</b>	<i>Recurrent Neural Network</i>
<b>CNN</b>	<i>Convolutional Neural Network</i>
<b>mRMR</b>	<i>Minimum Redundancy Maximum Relevance</i>
<b>IG</b>	<i>Information Gain</i>
<b>MARC</b>	<i>Multilingual Amazon Reviews Corpus</i>
<b>SGD</b>	<i>Stochastic Gradient Descent</i>
<b>BERT</b>	<i>Bidirectional Encoder Representations from Transformers</i>
<b>GPT-3</b>	<i>Generative Pre-Training Transformer 3</i>
<b>Seq2seq</b>	<i>Sequence-to-Sequence</i>
<b>GPUs</b>	<i>Graphic Processing Units</i>

# SUMÁRIO

<b>1</b>	<b>INTRODUÇÃO</b>	<b>11</b>
1.1	OBJETIVOS	12
1.2	ESTRUTURA DO TRABALHO	12
<b>2</b>	<b>ANÁLISE DE SENTIMENTO</b>	<b>13</b>
2.1	DEFINIÇÃO	13
2.2	ETAPAS DA ANÁLISE DE SENTIMENTO	15
<b>2.2.1</b>	<b>Coleta de dados</b>	<b>15</b>
<b>2.2.2</b>	<b>Pré-processamento dos dados</b>	<b>16</b>
<b>2.2.3</b>	<b>Extração de características</b>	<b>17</b>
<b>2.2.4</b>	<b>Seleção de características</b>	<b>19</b>
2.3	ABORDAGENS DE ANÁLISE DE SENTIMENTO	20
<b>2.3.1</b>	<b>Baseada em léxico</b>	<b>21</b>
<b>2.3.2</b>	<b>Baseada em aprendizado de máquina</b>	<b>23</b>
<b>2.3.3</b>	<b>Híbrida</b>	<b>24</b>
2.4	ARQUITETURA TRANSFORMER	25
2.5	MODELO BERT	28
<b>2.5.1</b>	<b>Pré-treinamento</b>	<b>28</b>
<b>2.5.2</b>	<b>Fine-tuning</b>	<b>29</b>
2.6	MEDIDAS DE DESEMPENHO	30
<b>3</b>	<b>ANÁLISE DE SENTIMENTO MULTI-IDIOMAS</b>	<b>33</b>
3.1	ABORDAGENS DE ANÁLISE DE SENTIMENTO MULTI-IDIOMAS	33
<b>3.1.1</b>	<b>Baseada em léxico</b>	<b>33</b>
<b>3.1.2</b>	<b>Baseada em corpus paralelos</b>	<b>34</b>
<b>3.1.3</b>	<b>Baseada em aprendizado de máquina</b>	<b>35</b>
<b>3.1.4</b>	<b>Baseada em tradução de máquina</b>	<b>36</b>
3.2	TRABALHOS RELACIONADOS	36
<b>3.2.1</b>	<b><i>A Multi-Layer Network for Aspect-Based Cross-Lingual Sentiment Classification</i></b>	<b>37</b>
<b>3.2.2</b>	<b><i>A character-based convolutional neural network for language-agnostic Twitter sentiment analysis</i></b>	<b>38</b>
<b>3.2.3</b>	<b><i>A comparative study of machine translation for multilingual sentence-level sentiment analysis</i></b>	<b>39</b>
<b>4</b>	<b>SISTEMA DE ANÁLISE DE SENTIMENTO MULTI-IDIOMAS</b>	<b>43</b>

4.1	METODOLOGIA . . . . .	43
4.1.1	<b>Tokenização . . . . .</b>	<b>44</b>
4.1.2	<b>Obtenção do modelo para a análise de sentimento . . . . .</b>	<b>45</b>
4.1.3	<b>Abordagens utilizadas . . . . .</b>	<b>46</b>
4.2	BASE DE DADOS . . . . .	48
4.3	RESULTADOS E DISCUSSÕES . . . . .	50
4.3.1	<b>Avaliação dos tradutores de máquina . . . . .</b>	<b>50</b>
4.3.2	<b>Comparação dos resultados da abordagem nativa e de tradução de máquina . . . . .</b>	<b>52</b>
4.3.3	<b>Impacto da polaridade no desempenho . . . . .</b>	<b>53</b>
5	CONSIDERAÇÕES FINAIS . . . . .	55
	<b>REFERÊNCIAS . . . . .</b>	<b>57</b>

# 1 INTRODUÇÃO

O avanço da tecnologia e o alto consumo de serviços digitais vêm transformando a forma de comunicação na sociedade (SATTAR *et al.*, 2021). Reuniões profissionais, conversas com amigos e avaliações de produtos são realizadas em ferramentas digitais, tornando-se esses grandes meio de expressão de sentimentos. Obter recomendações de compras ou opiniões de um produto, hoje é uma tarefa trivial, que pode ser realizada analisando as avaliações escritas por usuários em plataformas de compras. Entender o que seu cliente sente sobre seu produto é muito importante para as marcas. De modo a analisar os sentimentos expressos por textos, muitas tarefas de *Natural Language Processing* (NLP) foram criadas. Uma delas, a análise de sentimento, visa identificar os sentimentos expressos nos textos e sua polaridade, informando se o sentimento é positivo ou negativo (CAMBRIA, 2016).

Apesar da análise de sentimento ser uma tarefa de interesse global, a maioria das pesquisas e aplicações são realizadas para o idioma inglês (ARAÚJO; PEREIRA; BENEVENUTO, 2020). Um fator que dificulta a análise de sentimentos em idiomas de menor representatividade global é a falta de recursos léxicos de qualidade para esses idiomas. Muitos dos sistemas de análise de sentimento utilizam recursos léxicos famosos, como o *SentiWordNet*, que estão totalmente em inglês (DASHTIPOUR *et al.*, 2016). Dessa forma, aplicações que desejam realizar análise de sentimento em múltiplos idiomas podem necessitar de múltiplas abordagens, cada uma específica para um determinado idioma, chamadas abordagens nativas (ARAÚJO; PEREIRA; BENEVENUTO, 2020).

De modo a obter melhores resultados na análise de sentimento multi-idiomas, novas abordagens foram criadas, como a utilização de modelos de aprendizado de máquina e tradução de bases de dados (SATTAR *et al.*, 2021). A abordagem de tradução vem recebendo uma atenção especial dos pesquisadores, dada sua fácil aplicação e bons resultados obtidos. Essa abordagem visa utilizar apenas um método de análise de sentimento, geralmente para o idioma inglês, e textos dos demais idiomas são traduzidos para o inglês de modo a serem analisados. Essa tradução pode ser realizada manualmente ou com o auxílio de ferramentas de tradução de máquina.

Em conjunto com essas abordagens, novos modelos computacionais surgem para melhorar o desempenho de tarefas de NLP. Um deles é o modelo BERT, baseado na arquitetura *transformer*, que tem apresentado excelentes resultados, considerado estado da arte em representações de textos e em muitas tarefas de NLP (DEVLIN *et al.*, 2018).

Nesse contexto, esse trabalho visa avaliar se a abordagem de tradução de máquina é uma alternativa eficiente para solucionar o problema da análise de sentimento multi-idioma. Para isso é avaliado o desempenho da abordagem nativa e abordagem de tradução de máquina, utilizando o modelo BERT em ambas as abordagens.

## 1.1 OBJETIVOS

Este trabalho tem por objetivo avaliar diferentes abordagens de análise de sentimento para múltiplos idiomas utilizando o modelo BERT. Os objetivos específicos deste trabalho são:

1. Identificar as principais abordagens aplicadas no problema de análise de sentimento.
2. Obter uma base de dados textual para aplicação da análise de sentimento.
3. Identificar os modelos BERT mais utilizados para cada um dos idiomas a serem trabalhados.
4. Desenvolver e avaliar um sistema de análise de sentimento de textos em múltiplos idiomas utilizando o modelo BERT.

## 1.2 ESTRUTURA DO TRABALHO

Este trabalho está organizado da seguinte forma:

- O Capítulo 2 apresenta os conceitos de análise de sentimento e suas abordagens mais comuns. É apresentada a definição da arquitetura *transformer* e do modelo BERT, utilizado nesse trabalho. Por fim, são definidas as medidas de desempenho utilizadas na avaliação da análise de sentimento.
- O Capítulo 3 apresenta as abordagens de análise de sentimento multi-idiomas. São apresentados os trabalhos relacionados com análise de sentimento multi-idiomas e os resultados obtidos pelos autores.
- O Capítulo 4 descreve o sistema de análise de sentimento multi-idiomas proposto. É definido o método, a base de dados e os experimentos realizados. São apresentados os resultados e discussões.
- O Capítulo 5 apresenta as conclusões do trabalho e sugestões de trabalhos futuros.

## 2 ANÁLISE DE SENTIMENTO

Este capítulo tem por objetivo conceituar análise de sentimento, apresentar suas abordagens aplicadas em texto e medidas de desempenho utilizadas na avaliação. A Seção 2.1 apresenta a definição de análise de sentimento. A Seção 2.2 apresenta as principais etapas utilizadas na aplicação da análise de sentimento. A Seção 2.3 apresenta as principais abordagens de análise de sentimento encontradas na literatura. A Seção 2.4 apresenta a arquitetura *transformer*. A Seção 2.5 apresenta o modelo BERT. A Seção 2.6 apresenta as principais medidas de desempenho utilizadas na avaliação da análise de sentimento.

### 2.1 DEFINIÇÃO

Embora o estudo de NLP já venha sendo realizado há muitas décadas, foi somente a partir do século XXI que a análise sobre o sentimento expresso em textos virou uma área de estudo muito desejada por pesquisadores (LIU, 2012). Percebendo a grande capacidade de aplicação nas mais diversas áreas, criou-se uma indústria em torno dessa temática, o que motivou o crescimento desse estudo (Figura 1). Outro ponto que favoreceu seu desenvolvimento é a quantidade de informação disponíveis para serem utilizadas. A era digital possibilitou o armazenamento de um excesso de informações, principalmente com o crescimento desenfreado das redes sociais. As pessoas têm se expressado cada vez mais pela internet, comportamento que tornou a indústria da análise de sentimento uma área importante no entendimento do sentimento do público.

A análise de sentimento está diretamente relacionada com mineração de opiniões, extração de opiniões, mineração de sentimentos e computação afetiva. Apesar de tarefas diferentes, elas são todas pertinentes à mesma área de estudo (LIU, 2012). Os termos “sentimento” e “opinião” são constantemente encontrados na literatura referindo-se à mesma ideia. Diante disso, esse trabalho utilizará o termo “sentimento”.

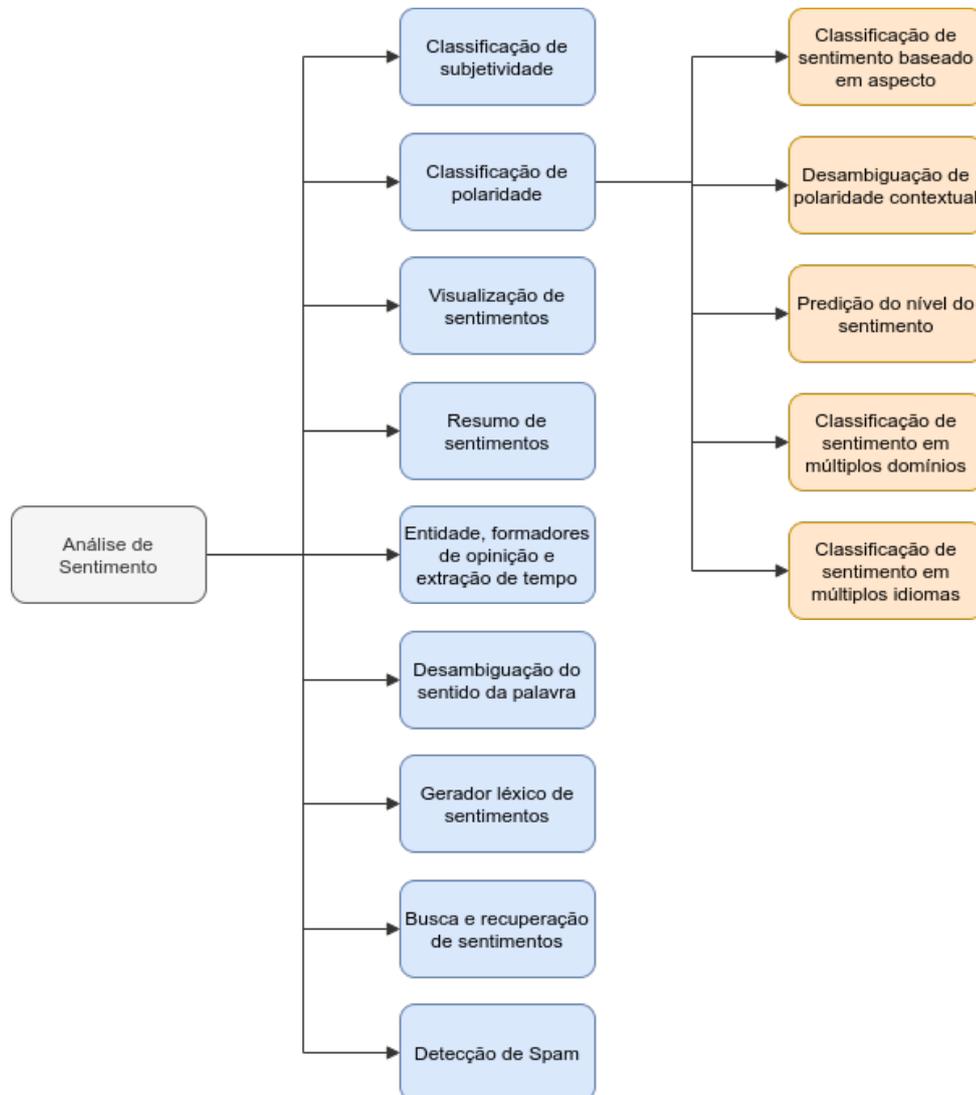
Figura 1 – Buscas por “*Sentiment Analysis*” no Google desde 2004



Fonte: <https://trends.google.com>

Longe de ser uma tarefa única e trivial, a análise de sentimento engloba muitos dos problemas que NLP tenta resolver (BIRJALI; KASRI; BENI-HSSANE, 2021). A Figura 2 ilustra algumas das tarefas aplicáveis em análise de sentimento. Esse trabalho foca na tarefa de “Classificação de polaridade”.

Figura 2 – Tarefas realizadas pela análise de sentimento



Fonte: Adaptado de Birjali, Kasri e Beni-Hssane (2021)

O processo de análise de sentimento pode ser realizado por meio das seguintes tarefas: identificar os sentimentos expressos em textos; detectar suas respectivas polaridades; e detectar as subjetividades (CAMBRIA, 2016). A tarefa de identificação de sentimentos é responsável por reconhecer quais os sentimentos estão expressos no conteúdo sendo analisado. Habitualmente, essa tarefa analisa os adjetivos e advérbios presentes no conteúdo. Por exemplo, na seguinte frase “essa música é linda”, é possível identificar o sentimento expresso pelo adjetivo “linda”. A detecção de polaridade visa categorizar a informação dada a sua orientação semântica, resultando se o sentimento é positivo ou negativo (CAMBRIA, 2016). A detecção de subjetividade é a tarefa responsável por verificar se a informação analisada contém viés de opinião (LO *et al.*,

2017). Informações subjetivas são frequentemente relacionadas a cultura de uma sociedade, não sendo possível utilizá-las como regra geral.

Liu (2012, p.4) define que a análise de sentimento pode ser aplicada em três níveis:

- **Documento:** é o nível mais simples de aplicação da análise de sentimento. Seu objetivo é identificar o sentimento e polaridade expressos pelo autor no documento (FELDMAN, 2013). Esse nível admite que o documento expressa um sentimento em relação a algum tópico específico, e não é aplicável em documentos que realizam avaliações ou comparações de vários itens ou tópicos.
- **Sentença:** caso o documento apresente vários sentimentos diferentes sobre o mesmo tópico, é necessária uma análise mais acurada sobre o texto. É nesse caso que pode ser aplicado o nível de sentença (FELDMAN, 2013). Ele visa identificar o sentimento e polaridade expressos em uma determinada sentença. Nesse nível além de positivo e negativo, o sentimento pode ser neutro, indicando não haver uma opinião expressa na sentença (LIU, 2012).
- **Entidade e Aspecto:** os níveis acima mencionados referem-se a documentos, ou sentenças, expondo sentimentos sobre uma única entidade, ou aspecto. Para atender casos com múltiplas entidades deve-se utilizar esse nível (FELDMAN, 2013). O nível de Entidade e Aspecto desconsidera as estruturas textuais e analisa diretamente o sentimento e seu alvo, seja ele uma entidade ou aspecto (característica) (LIU, 2012). Por exemplo, em *reviews* de produtos é comum encontrarmos uma análise sobre várias características do produto, algumas positivas, outras negativas. Nesse caso esse nível de análise de sentimento é aplicado.

## 2.2 ETAPAS DA ANÁLISE DE SENTIMENTO

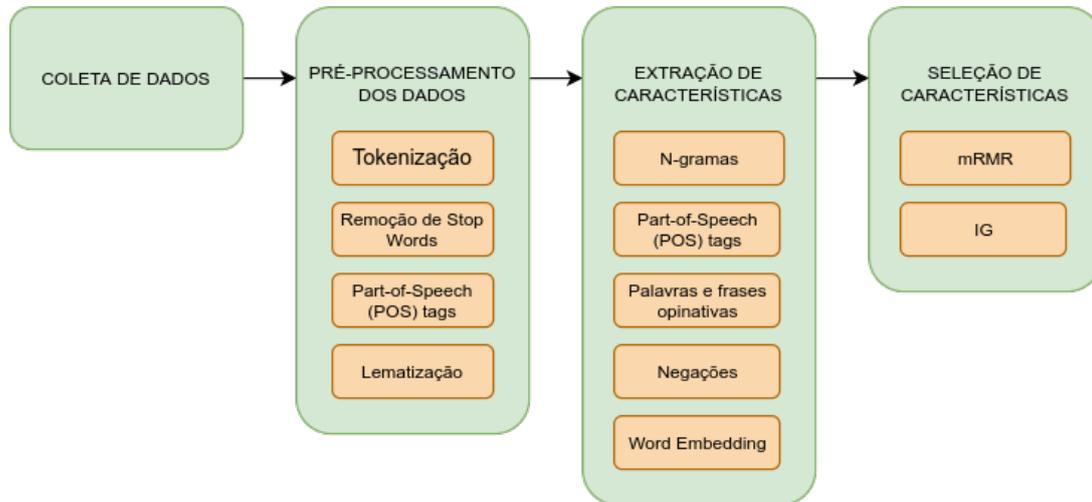
Birjali, Kasri e Beni-Hssane (2021, p.6) definem as principais etapas presentes na análise de sentimento como: coleta de dados, pré-processamento dos dados, extração de características e seleção de características (Figura 3).

### 2.2.1 Coleta de dados

A coleta de dados é a primeira tarefa para a realização da análise de sentimento. Mesmo aparentando ser uma tarefa trivial, a quantidade e qualidade dos dados coletados impacta diretamente no resultado da análise de sentimento. Por isso essa etapa deve ser realizada com muita cautela (BIRJALI; KASRI; BENI-HSSANE, 2021).

Os dados podem ser obtidos através da realização de pesquisas, bases de dados públicas e privadas ou consumidos de redes sociais e outras plataformas através de *Application Pro-*

Figura 3 – Etapas da análise de sentimento



Fonte: Adaptado de Birjali, Kasri e Beni-Hssane (2021)

*gramming Interface* (API). Dada a diversidade de origens dos dados, eles podem estar nos mais variados formatos, como TXT, CSV, XML ou JSON.

## 2.2.2 Pré-processamento dos dados

Geralmente, os dados coletados para a realização da análise de sentimento são não estruturados (BIRJALI; KASRI; BENI-HSSANE, 2021). Essa categoria de dados contém muito ruído e informações desnecessárias para sua análise. Por isso, a etapa de pré-processamento se torna importante.

Nessa etapa, o tratamento dos dados é realizado para remover ruídos, transformar os dados (para poderem ser utilizados em algoritmos de aprendizagem) e reduzir a dimensionalidade dos dados, simplificando a etapa de extração de características (ASGHAR *et al.*, 2014). No caso de remoção de ruídos, palavras que não causam impacto na polaridade do texto são removidas, como artigos, preposições, pontuações e caracteres especiais.

Birjali, Kasri e Beni-Hssane (2021, p.6) destacam as principais tarefas nessa etapa como:

- **Tokenização**: divisão do texto em elementos textuais menores (*tokens*) (BIRJALI; KASRI; BENI-HSSANE, 2021). Um texto é dividido em sentença, as sentenças são divididas em palavras.
- **Remoção de *Stop Words***: remoção de palavras que não contribuem em nada no resultado da análise de sentimento, como artigos e preposições (BIRJALI; KASRI; BENI-HSSANE, 2021).

- **Lematização:** conversão de palavras para sua forma normal, removendo assim tempos verbais, gerúndios e conjugações verbais (BIRJALI; KASRI; BENI-HSSANE, 2021). Verbos conjugados como “assistiu” e “ganhou”, passam para sua forma normal “assistir” e “ganhar”, por exemplo.

### 2.2.3 Extração de características

O objetivo dessa etapa é estimar características importantes que permitam determinar a polaridade da palavra, sentença ou documento analisado. É uma etapa muito importante e influencia diretamente no resultado da análise de sentimento.

As características utilizadas na análise de sentimento podem ser divididas em quatro categorias: sintática, semântica, *link-based* e estilística (ABBASI; CHEN; SALEM, 2008).

- **Sintáticas:** largamente aplicada em análise de sentimento, utilizam a função que cada termo desempenha na frase ou documento analisado, como verbos, substantivos, adjetivos e advérbios.
- **Semânticas:** visam classificar o sentimento utilizando regras gramaticais como sinônimos e antônimos.
- **Link-based:** canais de comunicação da internet relacionados tendem a apresentar o mesmo sentimento. Para isso são utilizados *links* e citações de documentos da internet para classificar outros documentos. Jornais, fóruns e *blogs* que apresentam elevado número de relações podem ter a mesma visão e sentimento sobre determinado assunto. Dada sua limitação, não é um tipo de característica muito utilizado.
- **Estilísticas:** utilizam informações léxicas e gramaticais na tentativa de identificação do sentimento, como o tamanho dos parágrafos, quantidade de letras por palavras ou quantidade de palavras por frases. Wiebe *et al.* (2004) identificou que textos subjetivos apresentam maior número de palavras únicas, quando comparados com textos objetivos. Contudo, não está clara a eficiência desse tipo de característica, por isso também não é muito utilizada na análise de sentimento.

Birjali, Kasri e Beni-Hssane (2021, p.7) destacam as características, sintáticas e semânticas, mais utilizadas na análise de sentimento:

- **N-gramas:** é uma sequência contínua de  $N$  termos em uma frase (DASHTIPOUR *et al.*, 2016). Casos em que  $N = 1$  são chamados unigramas, para  $N = 2$  bigramas, e  $n = 3$  trigramas. Sua principal característica é a consideração da ordem dos termos na frase, já que palavras como “não” e “nunca” podem alterar a polaridade do termo a seguir (NASSIR-TOUSSI *et al.*, 2014). Na frase “eu não venci a partida”, por exemplo, se utilizado  $N = 2$ ,

os bigramas serão: “eu não”, “não venci”, “venci a”, “a partida”. É possível identificar que mesmo a palavra “venci” indicando uma polaridade positiva, o termo anterior “não” altera essa polaridade. É uma característica muito utilizada com *Part-of-Speech*.

- ***Part-of-Speech (POS) tags***: são rótulos que identificam a classificação gramatical de uma palavra, como substantivos, adjetivos e verbos (BIRJALI; KASRI; BENI-HSSANE, 2021). Esse conhecimento é muito importante, pois os adjetivos e advérbios são fortes indicadores do sentimento expresso no texto. Textos de sentimento positivo tendem a apresentar mais advérbios superlativos, como “muito alto” e “altíssimo”, enquanto textos de sentimento negativo apresentam mais verbos de tempos passados (LO *et al.*, 2017). É uma característica muito utilizada em conjunto com n-gramas.
- **Palavras e frases opinativas**: são comumente utilizadas na expressão de sentimentos positivos e negativos (BIRJALI; KASRI; BENI-HSSANE, 2021). Como as palavras “ótimo” e “excelente” referindo-se aos sentimentos positivos, e “ruim” e “péssimo” referindo-se aos sentimentos negativos. Além da utilização de advérbios, também podem ser categorizados substantivos, verbos e expressões próprias do idioma trabalhado em questão.
- **Negações**: são palavras capazes de inverter a polaridade do sentimento. Sua grande maioria são termos negativos como “nenhum”, “ninguém” e “nunca”, porém nem sempre estão relacionados com sentimento negativo. Por exemplo, a frase “...ficou ferido” pode ser negativa, contudo a frase “ninguém ficou ferido” é positiva.

Diferentemente do processamento de áudio e vídeo, algumas abordagens de análise de sentimento são incapazes de tratar com dados em formato textual (SHI *et al.*, 2018). Por isso, há a necessidade de converter os textos para vetores numéricos, formato permitido por essas abordagens. *Word Embedding* é um modelo de conversão de textos em representações numéricas, mantendo seu valor semântico e sintático (GHANNAY *et al.*, 2016).

Os modelos de *Word Embedding* podem ser divididos em duas categorias, modelos baseados em arquitetura de contagem e baseados em arquitetura de predição (GANGADHARAN *et al.*, 2020). Os modelos baseados em arquitetura de contagem utilizam técnicas como *Count Vector* e *Term Frequency, Inverse Document Frequency (TF-IDF)*:

- **Count Vectors**: é a arquitetura mais simples para converter textos em vetores (GANGADHARAN *et al.*, 2020). Primeiramente são identificadas todas as palavras no texto, e armazenadas em um vetor de referência. Após outros vetores com o mesmo tamanho do de referência são criados para cada frase do texto, e nele são registradas as ocorrências de cada palavra na frase.
- **TF-IDF**: é a arquitetura mais comum para converter textos em vetores (GANGADHARAN *et al.*, 2020). Ela realiza a contagem de ocorrências de uma palavra, e normaliza com o

tamanho do texto analisado. Porém, um texto apresenta muitas *stop words*, que não são importantes. Para resolver isso é utilizado o fator de frequência inversa do documento. Assim termos com maior frequência são tratados com menor importância.

Os modelos baseados em arquitetura de predição utilizam *Continuous Bag of Words* (CBOW) e *Skip-Gram*.

- **CBOW**: tem o objetivo de prever uma palavra com base no contexto dado (GHANNAY *et al.*, 2016). Utiliza-se uma função logarítmica linear em que seu parâmetro de entrada é o texto. Essa função deve selecionar a palavra mais adequada ao contexto.
- **Skip-Gram**: similar a CBOW, visa prever o contexto com base em uma palavra (GHANNAY *et al.*, 2016). Também é utilizada uma função logarítmica linear onde o texto é o parâmetro de entrada.

A literatura cita os principais modelos de *Word Embedding* como Word2Vec, GloVe e fastText (BIRJALI; KASRI; BENI-HSSANE, 2021; UPRETI *et al.*, 2019; GHANNAY *et al.*, 2016).

## 2.2.4 Seleção de características

Na etapa de extração de características, muitas informações irrelevantes e redundantes são trazidas pelas características. De modo a considerar apenas as informações relevantes na análise de sentimento, é realizada a etapa de seleção de características (THAKKAR; CHAUDHARI, 2021). Essa etapa remove as informações irrelevantes de características extraídas, para reduzir a dimensão do conjunto de características, auxiliando assim o processo de classificação do sentimento. Através da redução da complexidade do classificador também é alcançada uma maior capacidade de generalização, o que é de grande vantagem para a reutilização do mesmo.

Agarwal e Mittal (2013, p.16) definem dois principais métodos utilizados na seleção de características para análise de sentimento:

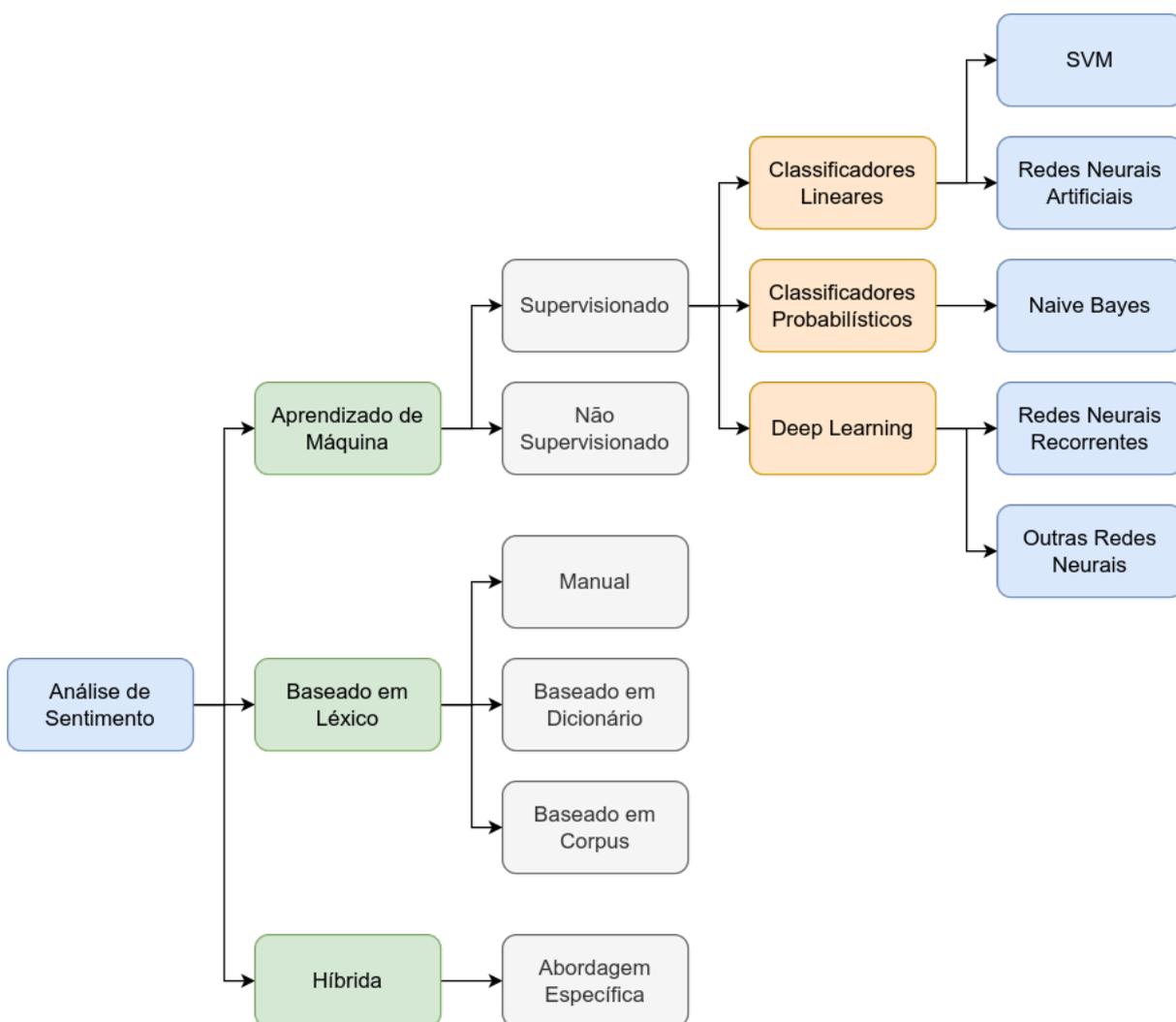
- **Minimum Redundancy Maximum Relevance (mRMR)**: é um método para identificar as características que melhor identificam uma classe do problema. São selecionadas as características que possuem maior relação com a classe e menor relação com outras características.
- **Information Gain (IG)**: é um método para identificar as características mais relevantes para determinada classe do problema. São calculados valores para as características individualmente.

## 2.3 ABORDAGENS DE ANÁLISE DE SENTIMENTO

A análise de sentimento é um campo de estudo em contínuo crescimento, aplicado nas mais diversas áreas, em que pesquisadores estão constantemente validando e comparando novas abordagens (BIRJALI; KASRI; BENI-HSSANE, 2021).

Os sistemas de análise de sentimento podem ser categorizados conforme a abordagem utilizada (Figura 4): baseada em léxico, aprendizado de máquina ou híbrida (BIRJALI; KASRI; BENI-HSSANE, 2021). A abordagem híbrida utiliza uma combinação das demais abordagens, visando alcançar a estabilidade da abordagem baseada em léxico e a generalização, e acurácia encontradas em abordagens de aprendizado de máquina. Nessas abordagens usualmente dicionários são utilizados como parâmetro de entrada no treinamento de modelos.

Figura 4 – Categorização das abordagens de análise de sentimento



Fonte: Adaptado de Taboada (2016).

### 2.3.1 Baseada em léxico

Também conhecida como *lexicon-based* ou *rule-based*, sugere que o sentimento geral de um texto seja a soma do sentimento de cada palavra ou frase nele presente (TABOADA, 2016). Por isso, essa abordagem é muito utilizada principalmente nos níveis de análise de sentença e entidade.

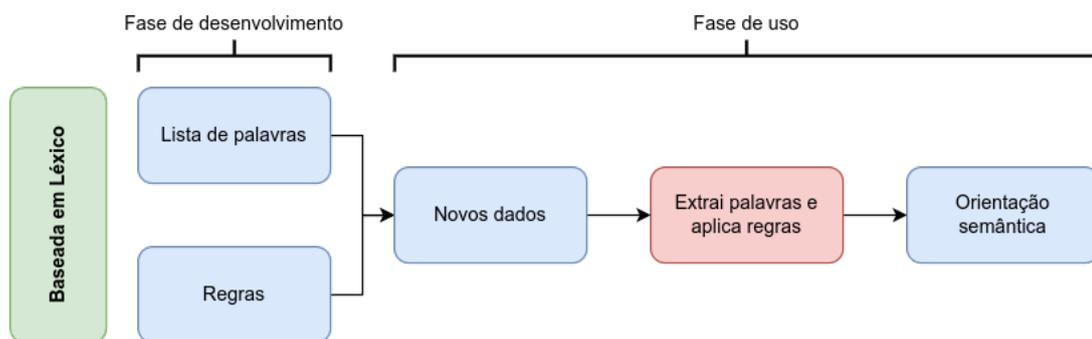
A identificação do sentimento é realizada utilizando um dicionário como referência, e aplicando algumas regras pré-definidas (CAMBRIA, 2016). Esse dicionário contém as palavras e suas respectivas polaridades, “ótimo” de polaridade positiva e “ruim” de polaridade negativa, por exemplo. Para a definição da polaridade é utilizada uma pontuação, podendo ser  $-1$ ,  $0$  e  $1$  para negativo, neutro e positivo respectivamente, ou pontuações maiores que representam a intensidade do sentimento (BIRJALI; KASRI; BENI-HSSANE, 2021). Alguns dicionários populares para realização da análise de sentimento no setor financeiro são *General Inquirer (GI)*, *Harvard IV-4 (HIV4)*, *Diction* e *Loughran and McDonald’s (LM)* (MISHEV *et al.*, 2020).

Essa abordagem é dividida em duas fases: desenvolvimento e uso (Figura 5) (TABOADA, 2016). A fase de desenvolvimento é onde ocorre a seleção ou construção do dicionário a ser utilizado. Birjali, Kasri e Beni-Hssane (2021) definem as três principais técnicas utilizadas para a construção do dicionário:

- **Manual:** a criação de todo o dicionário com as palavras e respectivas polaridades é realizada inteiramente por humanos. É um processo com ótimo resultado, porém é custoso financeiramente e consome muito tempo, por isso é normalmente utilizado em conjunto com as técnicas de dicionário ou corpus (MEDHAT; HASSAN; KORASHY, 2014). Recentemente alguns dicionários foram criados com essa técnica através de *crowdsourcing*, ou seja, colaboração pela internet de um grupo de pessoas com o mesmo objetivo.
- **Baseada em dicionário:** a principal ideia dessa técnica é que palavras sinônimas possuem a mesma polaridade, e antônimos polaridade oposta. A partir disso, o dicionário é criado com base em vários outros dicionários (BIRJALI; KASRI; BENI-HSSANE, 2021). Inicialmente um pequeno conjunto de palavras, com suas respectivas polaridades, é definido de forma manual. Então são utilizados algoritmos de busca, para encontrar sinônimos e antônimos dessas palavras em outros dicionários. Essa operação é realizada de forma recursiva, ou seja, para cada nova palavras inserida no dicionário. Ao final uma inspeção manual pode ser realizada, de modo a corrigir erros mais graves. Essa técnica é computacionalmente acessível e simples de ser aplicada para a criação de um dicionário necessário para o problema em questão.
- **Baseada em corpus:** sua principal ideia é a identificação de polaridade através de padrões sintáticos. Da mesma forma que a técnica baseada em dicionário, a baseada em corpus também precisa de um conjunto inicial de palavras com suas respectivas polaridades (BIR-

JALI; KASRI; BENI-HSSANE, 2021). A partir disso essa técnica utiliza padrões sintáticos e convenções linguísticas para definir a polaridade de novas palavras. Por exemplo, adjetivos interligados por conjunções, como "prático e barato", tendem a apresentar a mesma polaridade. Sua maior vantagem é a simplicidade de utilização, já que não precisa de vários dicionários diferentes, sendo tudo tratado no mesmo texto. Porém, necessita de um grande conjunto de dados iniciais para apresentar bons resultados.

Figura 5 – Fases da abordagem baseada em léxico



Fonte: Adaptado de Taboada (2016).

De posse do dicionário, a fase de uso visa aplicá-lo para a identificação da polaridade das palavras encontradas no texto analisado. Após encontrada a correspondência de uma palavra no dicionário, a polaridade da sentença ou do texto é definida conforme

$$Polaridade = \frac{Pos - Neg}{Pos + Neg}, \quad (2.1)$$

onde *Pos* é a quantidade de sentimentos positivos encontrados e *Neg* a quantidade de sentimentos negativos (MISHEV *et al.*, 2020). É o resultado dessa agregação de polaridades positivas e negativas que indica a orientação semântica de todo o texto. Para  $Polaridade > 0$  o texto é classificado como positivo, caso  $Polaridade < 0$  o texto é classificado como negativo.

Apesar de ser amplamente utilizada em análise de sentimento, esta abordagem apresenta algumas limitações. A primeira é a necessidade de um dicionário específico para a área e contexto do problema em questão (CAMBRIA, 2016). Como cada área possui termos técnicos específicos, o compartilhamento de dicionários entre área pode não representar corretamente o sentimento desejado. O contexto também pode apresentar divergências no sentimento, dado que uma característica pode ser positiva em determinado contexto, e negativa em outro (BIR-JALI; KASRI; BENI-HSSANE, 2021). Outra limitação é a quantidade de novas expressões sendo utilizadas, normalmente em rede sociais e em contextos diferentes. Novas formas de expressão como abreviaturas e *emojis* surgem com frequência, o que impossibilita manter um dicionário sempre atualizado.

### 2.3.2 Baseada em aprendizado de máquina

As abordagens baseadas em aprendizado de máquina utilizam características sintáticas e linguísticas para determinar o sentimento de um texto (MEDHAT; HASSAN; KORASHY, 2014).

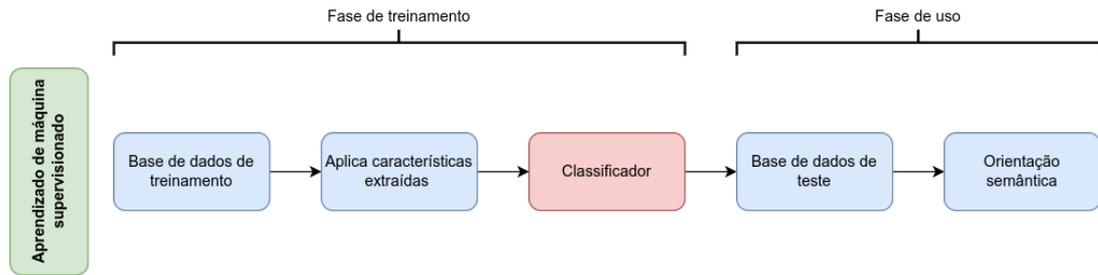
A grande vantagem de utilizar uma abordagem de aprendizado de máquina é a capacidade de aprender de forma automática as características do sentimento a ser estimado, possuindo assim maior capacidade de generalização (MADHOUSHI; HAMDAN; ZAINUDIN, 2015).

De acordo com Medhat, Hassan e Korashy (2014), os algoritmos de aprendizado de máquina podem ser divididos em dois grupos principais:

- **Aprendizado supervisionado:** um conjunto de dados é utilizado para treinar um algoritmo a distinguir o sentimento (TABOADA, 2016). No caso da análise de sentimento, a base de dados deve possuir a respectiva polaridade, positiva, neutra ou negativa (MEDHAT; HASSAN; KORASHY, 2014). Esse modelo é aplicado em duas fases distintas: treinamento e uso (Figura 6) (TABOADA, 2016). Na fase de treinamento, é utilizada a base de dados rotulada justamente para treinar o modelo a identificar a polaridade do sentimento com base nas características extraídas do texto. De posse do classificador treinado, utiliza-se a base de dados de testes para a aplicação do classificador. Contudo, é difícil criar as bases de dados rotuladas corretamente para cada problema (BIRJALI; KASRI; BENI-HSSANE, 2021). Um dos problemas é que os dados devem ser rotulados manualmente, tornando a tarefa trabalhosa e passível de subjetividade do rotulador.
- **Aprendizado não-supervisionado:** modelos não supervisionados não necessitam de dados de treinamento rotulados. As características extraídas nas etapas anteriores a classificação são as principais informações utilizadas por estes modelos (MEDHAT; HASSAN; KORASHY, 2014). Esses modelos normalmente utilizam técnicas de agrupamento, separando os dados em diferentes categorias, porém sem especificar quais, sendo essa uma das desvantagens desse modelo. Eles são atrativos, pois, não necessitando de dados de treinamento rotulados, a coleta de dados não estruturados, principalmente textos, pode ser realizada facilmente. Entretanto, tendem a ser mais complexos no entendimento e dificultam sua avaliação de desempenho, dado que não apresentam explicitamente as polaridades classificadas (BIRJALI; KASRI; BENI-HSSANE, 2021).

A maioria das abordagens de análise de sentimento utilizam o aprendizado supervisionado. Dentre os trabalhos analisados, os classificadores mais utilizados são *Naïve Bayes*, *Support Vector Machine (SVM)*, *Artificial Neural Network (ANN)* e *Recurrent Neural Network (RNN)*.

Figura 6 – Fases da abordagem de aprendizado de máquina supervisionado



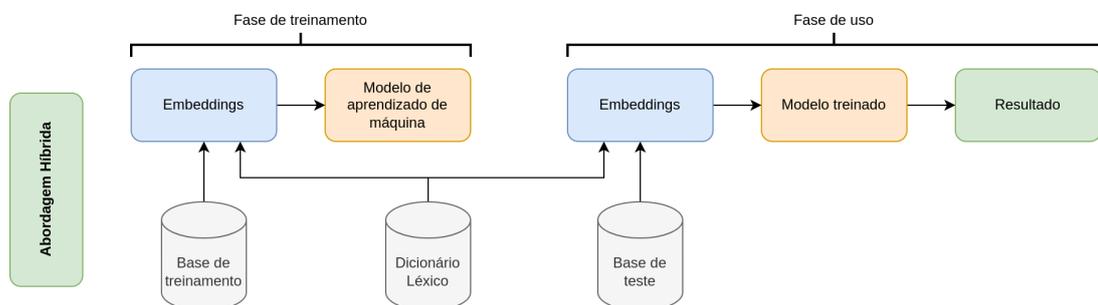
Fonte: Adaptado de Taboada (2016).

### 2.3.3 Híbrida

As abordagens de análise de sentimento híbridas utilizam dicionários léxicos em conjunto com modelos de aprendizado de máquina (BIRJALI; KASRI; BENI-HSSANE, 2021). A combinação dessas abordagens visa obter melhores resultados na análise de sentimento, beneficiando-se de toda informação presente nos dicionários léxicos e da flexibilidade dos modelos de aprendizado de máquina. Nessa abordagem, os dicionários léxicos são utilizados na construção das representações do texto (*embedding*), enquanto o modelo de aprendizado de máquina é utilizado para classificar os dados.

Essa abordagem também pode ser dividida em duas etapas: treinamento e uso. Na fase de treinamento, os *embeddings* são gerados a partir do dicionário léxico e da base de treinamento. Esses *embeddings* são utilizados como parâmetro de entrada do modelo de aprendizado de máquina, de modo a realizar o seu treinamento. Na fase de uso, os *embeddings* são gerados a partir do dicionário léxico e da base de teste. Após, os *embeddings* são passados como parâmetro de entrada para o modelo já treinado, de modo a obter o resultado da análise de sentimento (Figura 7).

Figura 7 – Diagrama da abordagem híbrida



Fonte: Adaptado de Taboada (2016).

## 2.4 ARQUITETURA TRANSFORMER

Em 2017, visando obter melhores resultados nas tarefas de processamento de linguagem natural, pesquisadores apresentaram uma nova arquitetura chamada *transformer* (VASWANI *et al.*, 2017). Essa arquitetura é baseada em *encoder-decoder* e implementa um mecanismo de auto-atenção, capaz de aprender, dado um contexto, a relação entre os dados de entrada. Originalmente proposto para problemas *Sequence-to-Sequence* (Seq2seq), o *transformer* é atualmente utilizado nos mais diversos problemas, sendo a arquitetura base para modelos de aprendizado de máquina como o *Generative Pre-Training Transformer 3* (GPT-3) e o BERT. Apresentando desempenho superior a outras arquiteturas como RNN e *Convolutional Neural Network* (CNN), e alta capacidade de paralelização do processamento, o *transformer* é considerado a arquitetura mais avançada para representação de textos e é muito utilizada em tarefas de NLP, como em tradução de textos e análise de sentimento.

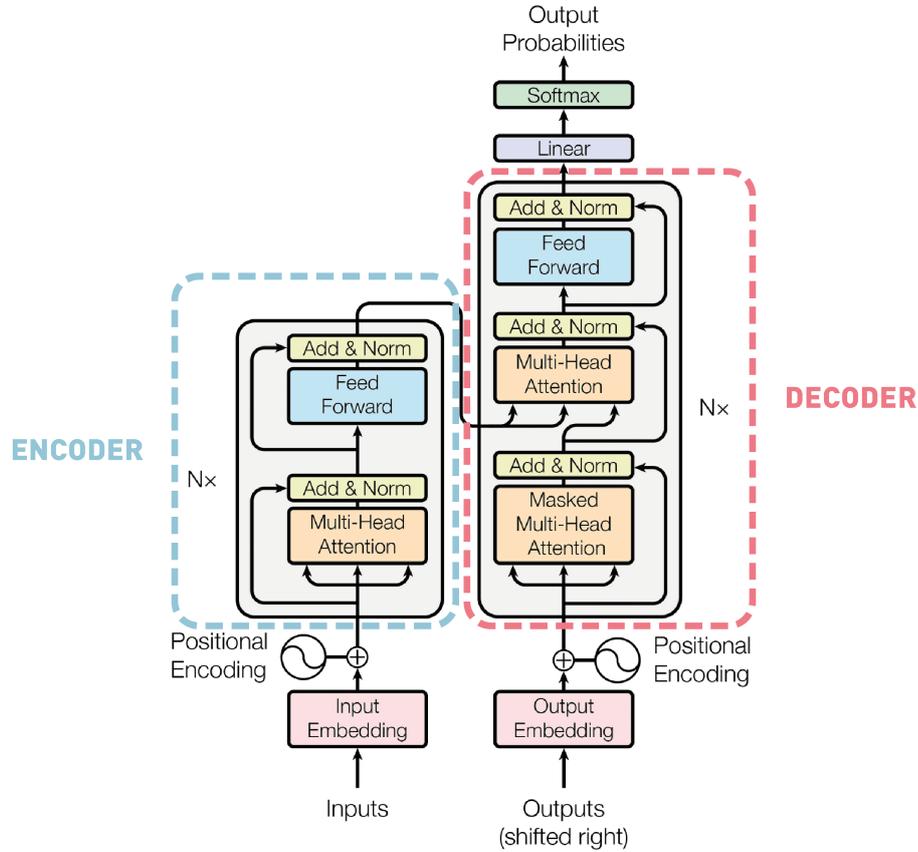
A arquitetura *transformer* é composta por dois principais componentes: *encoder* e *decoder* (Figura 8) (VASWANI *et al.*, 2017). O *encoder* é responsável por mapear uma sequência de *tokens*  $(x_1, \dots, x_n)$  para uma sequência de representações contínuas  $z = (z_1, \dots, z_n)$ . Para isso é implementado um mecanismo de atenção e uma rede *feed-forward*. O *decoder* é responsável por, dada a entrada  $z$ , gerar a sequência de saída  $(y_1, \dots, y_n)$ , gerando uma representação por vez. Cada processo do *decoder* recebe como entrada o resultado do *encoder* e o resultado do processo anterior do *decoder*, gerando um modelo auto-regressivo. Dessa forma, o *decoder* implementa uma rede *feed-forward* e dois mecanismos de atenção: um para a entrada proveniente do *encoder*, e outro para a entrada do processo anterior do *decoder*. O *transformer* implementa pilhas de *encoders* e *decoders*. Vaswani *et al.* (2017) definem a pilha de *encoders* composta de seis codificadores, cada um deles composto por um mecanismo de auto-atenção (*Multi-Head Attention*) e uma rede *feed-forward*. A pilha de *decoders* é composta por seis decodificadores, cada um deles composto por dois mecanismos de auto-atenção e uma rede *feed-forward*.

A arquitetura *transformer* não utiliza recorrência ou convoluções, mas implementa um mecanismo de autoatenção que realiza a interconexão de seus dois principais componentes *encoder* e *decoder* (VASWANI *et al.*, 2017). Esse mecanismo aplica uma função de atenção, permitindo que, ao gerar uma representação dos dados, sejam processadas as partes mais importantes dos dados de entrada de forma mais relevante que as partes menos importantes (BAHDANAU; CHO; BENGIO, 2014). A função de atenção no *transformer* é chamada de *Scaled dot-product attention* (Figura 9a), e é calculada por:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2.2)$$

As variáveis de entrada são as matrizes *query*  $Q$ , *key*  $K$  e *value*  $V$ , de dimensões  $d_k$ ,  $d_k$  e  $d_v$ , respectivamente (VASWANI *et al.*, 2017). A matriz  $Q$  contém todas as representações de uma palavra da frase, a matriz  $K$  contém todas as representações de todas as palavras da frase, e a matriz  $V$  representa a soma ponderada dos valores obtidos do *encoder*. O resultado é uma

Figura 8 – Arquitetura transformer



Fonte: Adaptado de Vaswani *et al.* (2017).

matriz contendo a representação de cada dado de entrada, baseada na relação desse dado com os demais.

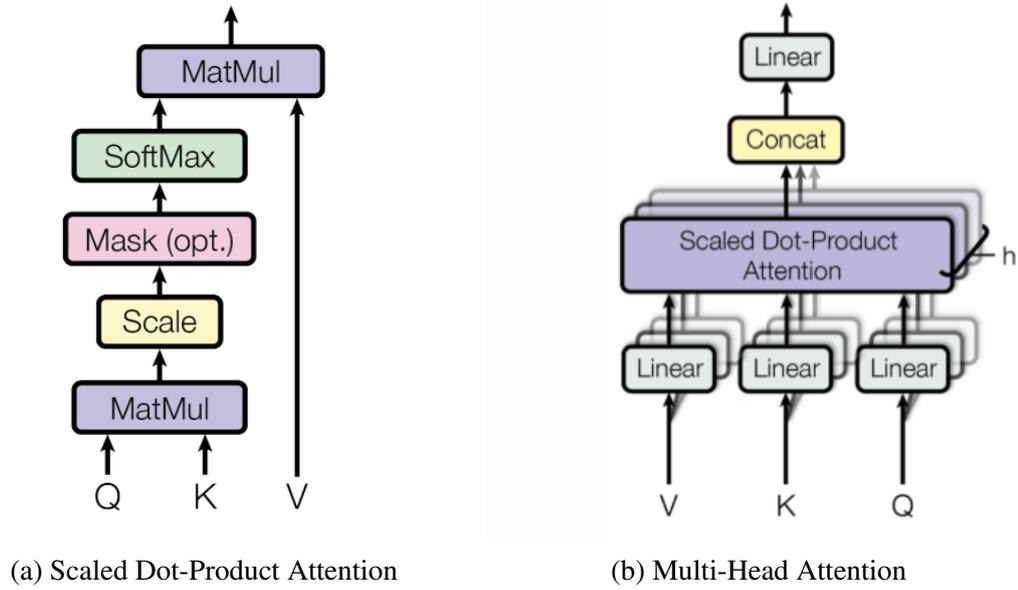
A execução da função de atenção uma única vez, ao gerar uma representação, pode resultar em relações enganosas entre os dados. Para evitar isso, o *transformer* aplica a função de atenção diversas vezes para um mesmo dado, e sua representação final é uma projeção linear de todos os resultados concatenados (Figura 9b) (VASWANI *et al.*, 2017). As matrizes  $Q$ ,  $K$  e  $V$  são linearmente projetadas utilizando matrizes de projeções diferentes, permitindo assim cada núcleo de atenção focar em diferentes posições dos dados de entrada. Todas as projeções podem ser calculadas paralelamente, em que Vaswani *et al.* (2017) utilizaram o poder computacional de *Graphic Processing Units* (GPUs) para paralelizar o processamento. Esse processo, chamado *Multi-Head Attention*, pode ser matematicamente descrito como:

$$MultiHead(Q, K, V) = Concat(head_1, \dots, head_h)W^O \quad (2.3)$$

onde  $head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$ ,

em que *head* são os núcleos de atenção,  $h$  é o número de núcleos de atenção e  $W$  são as diferentes matrizes de projeção.

Figura 9 – Mecanismo de atenção na arquitetura *transformer*



(a) Scaled Dot-Product Attention

(b) Multi-Head Attention

Fonte: Vaswani *et al.* (2017)

Dado que a arquitetura do *transformer* não utiliza recorrência ou convolução, de modo a permitir que o mecanismo de atenção seja aplicado corretamente, é preciso que as representações geradas dos dados de entrada contêm informações suficientes sobre a posição desse dado na sequência de entrada (VASWANI *et al.*, 2017). Para isso é gerado um vetor — *positional encoding* (*PE*) — que contém a informação sobre a posição do *token*. Esse vetor tem a mesma dimensão do vetor de *embeddings*, assim é possível agrupá-los. Vaswani *et al.* (2017) utilizaram as funções de seno e cosseno para gerar o vetor de posições:

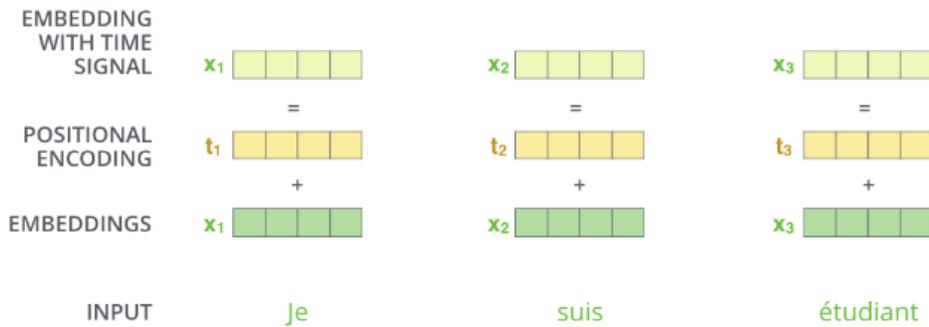
$$\begin{aligned} PE_{(pos,2i)} &= \sin(pos/10000^{2i/d_{model}}) \\ PE_{(pos,2i+1)} &= \cos(pos/10000^{2i/d_{model}}), \end{aligned} \quad (2.4)$$

onde *pos* é a posição do *token* processado e *i* a dimensão do vetor. O resultado desse componente é um vetor de *embeddings* contendo informações sobre o *token* e sua posição, como pode ser visualizado na Figura 10.

Os componentes de *encoder* e *decoder* implementam, ao final de seu fluxo, uma rede *feed-forward* (VASWANI *et al.*, 2017). Essa rede é responsável por sintetizar os vetores de atenção resultantes das camadas de atenção para uma forma possível de ser reinserida no fluxo. A rede *feed-forward* implementa duas transformações lineares, e a função de ativação ReLU, e pode ser descrita matematicamente por:

$$FFN(x) = \max(0, xW_1 + b_1)W_2 + b_2 \quad (2.5)$$

Figura 10 – Processo de aplicação do vetor de posições



Fonte: Vaswani *et al.* (2017)

## 2.5 MODELO BERT

BERT é um modelo de redes neurais profundas designado para pré-treinar representações bidirecionais de textos, considerado estado da arte para tarefas de NLP (DEVLIN *et al.*, 2018). É um modelo baseado na arquitetura *transformer*, porém como seu objetivo é criar representações de textos, ele utiliza apenas o componente de *encoder* dessa arquitetura. Sua maior característica é a utilização de mecanismos de auto-atenção bidirecionais, permitindo que ambas sequências do texto (esquerda e direita) sejam utilizadas como contexto ao gerar a representação de uma palavra.

Devlin *et al.* (2018) desenvolveram dois modelos em dois tamanhos: BERT<sub>BASE</sub> e BERT<sub>LARGE</sub>. Eles são diferenciados pelo número de *encoders*, número de núcleos de atenção, número de camadas ocultas e número de hiperparâmetros utilizados. O modelo BERT<sub>BASE</sub> é composto por 12 camadas de *encoders*, 768 camadas ocultas, 12 núcleos de atenção e 110 milhões de hiperparâmetros. O modelo BERT<sub>LARGE</sub> é composto por 24 camadas de *encoders*, 1024 camadas ocultas, 16 núcleos de atenção e 340 milhões de hiperparâmetros.

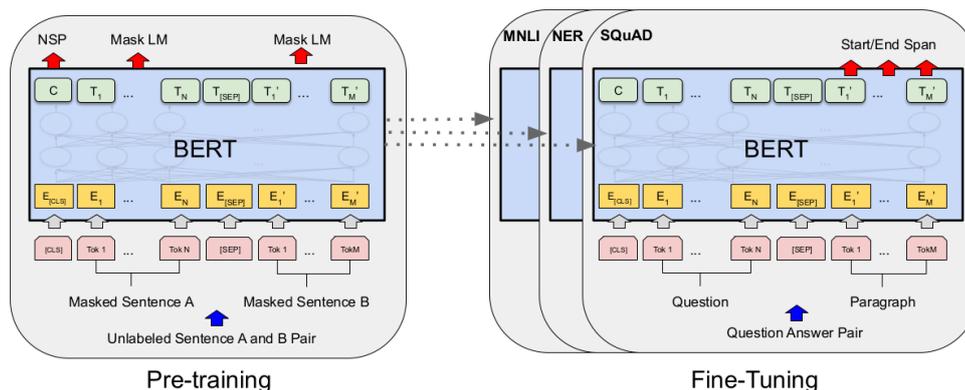
O modelo BERT é treinado em duas etapas distintas: pré-treinamento e *fine-tuning* (Figura 11) (DEVLIN *et al.*, 2018).

### 2.5.1 Pré-treinamento

A etapa de pré-treinamento utiliza uma grande quantidade de textos não rotulados para estimar os hiperparâmetros do modelo (DEVLIN *et al.*, 2018). Essa etapa tende a ser demorada e de alta demanda computacional, dessa forma é apenas executada para gerar um modelo base, que deve passar pelo *fine-tuning* para ser aplicado. Na geração dos modelos BERT<sub>BASE</sub> e BERT<sub>LARGE</sub>, foram utilizadas as bases de dados *BooksCorpus*, com 800 milhões de palavras, e a *English Wikipedia*, com 2.5 bilhões de palavras.

De modo a gerar sua característica de processamento bidirecional, não são utilizados modelos de linguagem *left-to-right* ou *right-to-left* no pré-treinamento, mas sim duas tarefas de

Figura 11 – Etapas de pré-treinamento e *fine-tuning* do modelo BERT



Fonte: Devlin *et al.* (2018)

aprendizado não supervisionado: *Masked LM* e *Next Sentence Prediction* (NSP).

- **Masked LM:** são mascarados, de forma aleatória, alguns *tokens* de entrada (DEVLIN *et al.*, 2018). Esse processo é realizado substituindo o *token* de entrada pelo *token* [MASK]. Após, o modelo trabalha na predição desse *token* mascarado, baseando-se no seu contexto a esquerda e direita. Na geração dos modelos BERT<sub>BASE</sub> e BERT<sub>LARGE</sub> foram mascarados 15% dos *tokens* de entrada de cada sentença. Dado que o *token* [MASK] apenas existe nessa etapa de pré-treinamento, utilizar esse *token* pode causar incompatibilidade na etapa de *fine-tuning*, que desconhece o *token* [MASK]. Dessa forma, o *token* [MASK] é utilizado em 80% dos casos, em 10% é utilizado outro *token* presente na sentença, e em 10% o *token* não é alterado.
- **Next Sentence Prediction:** tem por objetivo treinar o modelo para entender a relação entre duas sentenças do texto. Dessa forma, o modelo pode utilizar informações externas a sentença em processamento para gerar sua representação (DEVLIN *et al.*, 2018). Essa tarefa recebe como entrada duas sentenças do texto (A, B), e o modelo deve predizer se a sentença B é uma sequência direta de A ou uma sentença aleatória do texto. Devlin *et al.* (2018), ao escolherem os pares de sentenças, 50% das vezes foram escolhidas sentenças em sequência, e 50% das vezes sequências aleatórias do texto.

## 2.5.2 Fine-tuning

A etapa de *fine-tuning* tem como objetivo preparar um modelo pré-treinado para aplicá-lo em uma tarefa de NLP, como *question answering* ou análise de sentimento (DEVLIN *et al.*, 2018). Dada sua arquitetura baseada em *transformers*, a etapa de *fine-tuning* torna-se simples de ser implementada, em que apenas as entradas e saídas precisam ser adaptadas de forma apropriada. Assim, é significativamente mais rápida de ser aplicada. Devlin *et al.* (2018) finalizaram o *fine-tuning* de seus modelos em algumas horas quando utilizaram GPUs.

O treinamento nessa etapa é supervisionado, sendo utilizadas as bases de dados rotuladas para a tarefa específica. Essas bases são inseridas no modelo BERT para realizar o *fine-tuning* do modelo (DEVLIN *et al.*, 2018) para a tarefa desejada. De forma análoga às sentenças A e B utilizadas na etapa de pré-treinamento, a sentença e sua polaridade são utilizadas para treinar suas relações no *fine-tuning* para a tarefa de análise de sentimento. Uma nova camada é inserida após a última camada oculta do modelo, sendo essa responsável pela classificação dos dados para a tarefa aplicada.

Essa etapa utiliza *transfer-learning*, uma técnica que permite que modelos pré-treinados sejam reaproveitados para outras aplicações (OLIVAS *et al.*, 2009). Sua premissa é de que utilizar modelos pré-treinados com uma grande quantidade de dados como ponto de partida para resolver uma tarefa é mais vantajoso que realizar o treinamento de um novo modelo. Assim, é possível se beneficiar de todo treinamento não supervisionado já realizado em um modelo BERT, e fazer o ajuste fino para a tarefa desejada.

## 2.6 MEDIDAS DE DESEMPENHO

De modo a selecionar o melhor classificador para um problema específico, é preciso avaliar o seu desempenho. Para isso diversas medidas de desempenho podem ser utilizadas (BIRJALI; KASRI; BENI-HSSANE, 2021). A escolha das medidas utilizadas afeta diretamente como o classificador será mensurado e comparado entre os demais. As medidas de desempenho mais utilizadas na análise de sentimento são acurácia, precisão, *recall* e *F1-Score*. Para entender essas métricas é preciso antes o entendimento sobre matriz de confusão.

A matriz de confusão em problemas binários é uma matriz  $2 \times 2$  (Figura 12), que contabiliza a quantidade de amostras de uma base de dados classificadas corretamente ou não em problemas de classificação binária (SHOBHA; RANGASWAMY, 2018). Ela possui em sua diagonal principal as classificações corretas, verdadeiro positivo (VP) e verdadeiro negativo (VN), e na diagonal inversa os erros de classificação, falso positivo (FP) e falso negativo (FN) (CHICCO; JURMAN, 2020), em que:

- **VP**: amostras positivas que foram classificadas de forma positiva;
- **VN**: amostras negativas que foram classificadas de forma negativa;
- **FP**: amostras negativas que foram classificadas de forma positiva;
- **FN**: amostras positivas que foram classificadas de forma negativa;

Um classificador perfeito deve, ao final da classificação de todas as amostras de uma base de dados, computar zero registros falso positivos  $FP = 0$ , e zero falso negativos  $FN = 0$  (BIRJALI; KASRI; BENI-HSSANE, 2021).

Figura 12 – Matriz de confusão para problemas de 2 classes

		Valor Atual	
		Positivo	Negativo
Valor Previsto	Positivo	VP (Verdadeiro Positivo)	FP (Falso Positivo)
	Negativo	FN (Falso Negativo)	VN (Verdadeiro Negativo)

Fonte: Adaptado de Araújo, Pereira e Benevenuto (2020).

A partir da matriz de confusão pode-se estimar as seguintes medidas de desempenho:

- **Acurácia** ( $Acc$ ): é definida como a relação entre as classificações corretas e todas as classificações realizadas (CHICCO; JURMAN, 2020), conforme

$$Acc = \frac{VP + VN}{VP + VN + FP + FN}. \quad (2.6)$$

Seu resultado é um valor no intervalo fechado de 0 e 1, sendo 0 o pior cenário, com todas classificações erradas, e 1 o melhor cenário com todas classificações corretas.

Essa medida funciona muito bem quando as classes do problema estão balanceadas, ou seja, nenhuma classe possui um número consideravelmente maior de amostras que outra (CHICCO; JURMAN, 2020), sendo a medida de desempenho mais utilizada em problemas de classificação binária (SHOBHA; RANGASWAMY, 2018). Contudo, quando as classes não estão balanceadas, a acurácia não consegue medir a efetividade do classificador corretamente, podendo comprometer a análise se for a única medida a ser considerada.

De posse da acurácia, também é possível descobrir o erro de um classificador (BIRJALI; KASRI; BENI-HSSANE, 2021), definido por

$$E = 1 - Acc. \quad (2.7)$$

- **Precisão** ( $P$ ): definida como a razão entre as amostras classificadas corretamente como positivas sobre todas as classificações positivas, conforme

$$P = \frac{VP}{VP + FP}. \quad (2.8)$$

É uma métrica muito utilizada para analisar problemas em que a confiabilidade da classificação precisa ser alta (BIRJALI; KASRI; BENI-HSSANE, 2021).

- **Recall (R)**: definida como a razão entre as amostras classificadas corretamente como positivas sobre todas as amostras positivas, conforme

$$R = \frac{VP}{VP + FN}. \quad (2.9)$$

É utilizada em casos em que uma classe deve ser dominante para o classificador (BIRJALI; KASRI; BENI-HSSANE, 2021).

Ambas as métricas de precisão e *recall* não consideram o número de verdadeiros negativos (VN) (SHOBHA; RANGASWAMY, 2018). Por isso é possível que, mesmo um classificador acertando 99% das tentativas, sua precisão ou *recall* seja de 50%, por exemplo.

- **F1-Score**: é definida como a média harmônica entre as medidas de precisão e *recall* (CHICCO; JURMAN, 2020). Dessa forma, é muito utilizada na comparação de classificadores com alta precisão e baixo *recall*, ou vice-versa (BIRJALI; KASRI; BENI-HSSANE, 2021). Sendo uma métrica originária da matriz de confusão, o F1-Score é calculado por:

$$F1 = \frac{2 * VP}{2 * VP + FP + FN} = \frac{2 * P * R}{P + R}, \quad (2.10)$$

em que  $P$  é precisão e  $R$  o *recall*. Seu resultado é um valor no intervalo fechado de 0 e 1. O pior cenário ( $F1 = 0$ ) ocorre quando  $VP = 0$ , ou seja, quando todas as amostras positivas são classificadas como negativas. Já o melhor cenário ( $F1 = 1$ ), ocorre quando  $FN = FP = 0$ , ou seja, nenhuma classificação foi realizada de forma errônea. O *F1 - Score* se diferencia das outras medidas de desempenho por não contabilizar as classificações negativas corretas (TN), e por não ser simétrico entre as classes do problema. Ou seja, o valor  $F1$  para a classe positiva não necessariamente será igual ao  $F1$  da classe negativa (CHICCO; JURMAN, 2020).

- **Matthews Correlation Coefficient (MCC)**: é definido como o coeficiente de correlação entre os valores reais e os valores previstos pelo classificador (CHICCO; JURMAN, 2020). É uma medida que não sofre o impacto de bases dados não balanceadas, sendo muito utilizada para esses casos. Sua equação é definida por:

$$MCC = \frac{VP * VN - FP * FN}{\sqrt{(VP + FP)(FN + VN)(FP + VN)(VP + FN)}}. \quad (2.11)$$

Seu resultado é um valor no intervalo fechado de [-1, 1], em que  $MCC = -1$  representa um classificador que errou todas as classificações, e  $MCC = 1$  representa um classificador perfeito. O  $MCC$  é a única medida de desempenho para problemas de classificação binária que obriga o classificador a ter bons resultados tanto nas amostras positivas (TP) quanto nas amostras negativas (TN) (CHICCO; JURMAN, 2020).

### 3 ANÁLISE DE SENTIMENTO MULTI-IDIOMAS

Este capítulo tem por objetivo apresentar as abordagens utilizadas em análise de sentimento multi-idiomas e trabalhos relacionados. A Seção 3.1 apresenta as abordagens de análise de sentimento multi-idiomas mais utilizadas na literatura. A Seção 3.2 apresenta trabalhos relacionados com análise de sentimento multi-idiomas.

#### 3.1 ABORDAGENS DE ANÁLISE DE SENTIMENTO MULTI-IDIOMAS

As abordagens de análise de sentimento multi-idiomas são adaptações de abordagens que já se mostraram eficientes para resolver o problema de análise de sentimento com idioma único. Em especial abordagens para o idioma inglês, pois possui os melhores recursos léxicos, possibilitando um melhor desempenho na análise de sentimento (ARAÚJO; PEREIRA; BENEVENUTO, 2020). São combinadas várias técnicas utilizadas na análise de sentimento de um único idioma para dar origem a uma nova abordagem multi-idiomas.

Entre as abordagens mais comumente encontradas na literatura estão a baseada em léxico, baseada em corpus paralelos, baseada em aprendizado de máquina e baseada em tradução de máquina.

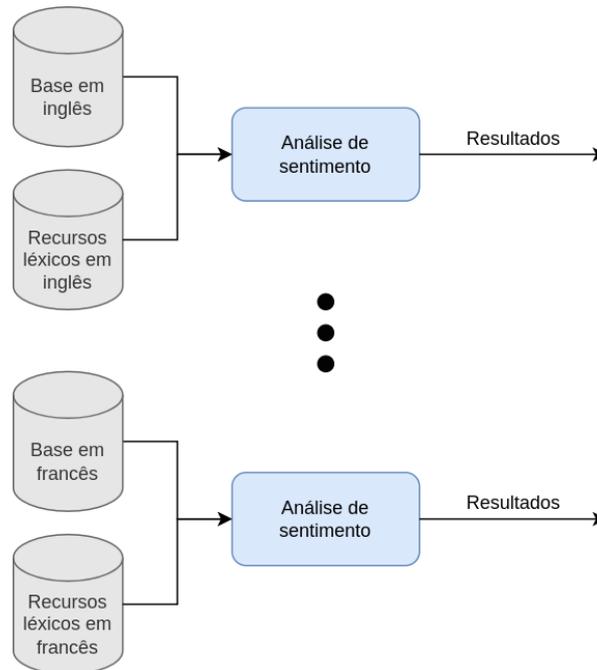
##### 3.1.1 Baseada em léxico

Essa abordagem consiste em realizar a análise de sentimento consultando um dicionário léxico específico para o idioma em questão. Cada idioma a ser analisado deverá possuir seu dicionário léxico (ARAÚJO; PEREIRA; BENEVENUTO, 2020). A criação desse dicionário pode ser realizada de forma manual ou com auxílio de tradução de máquina. Porém, nessa abordagem a tradução é realizada apenas no dicionário, e não no texto a ser analisado.

É uma abordagem manual, que considera cada idioma de forma independente, como ilustrado na Figura 13. Essa característica pode dificultar sua implementação, pois muitos idiomas carecem de bons recursos léxicos. Os dicionários existentes para os idiomas menos comuns podem não ser suficientes para a obtenção de um bom resultado na análise de sentimento, e construir um dicionário léxico não é uma tarefa trivial (LO *et al.*, 2017).

Conseqüentemente, o desempenho dessa abordagem está diretamente relacionado com a qualidade do dicionário léxico utilizado. Idiomas com escassez de recursos léxicos são prejudicados, nesses casos, não sendo essa a melhor opção.

Figura 13 – Abordagem multi-idioma baseada em léxico



Fonte: Elaborado pelo autor (2022).

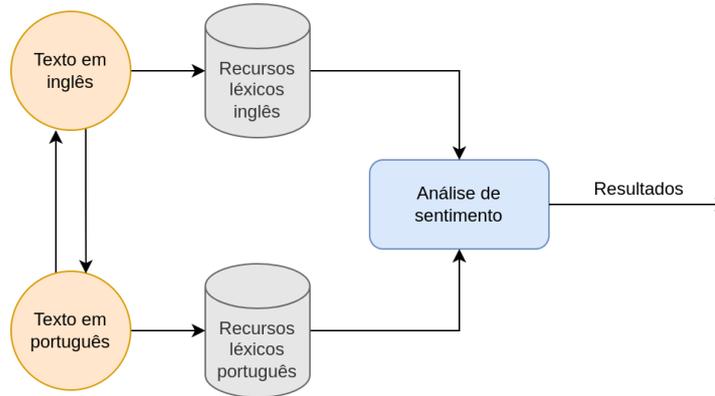
### 3.1.2 Baseada em corpus paralelos

Essa abordagem utiliza dois textos de mesmo valor semântico, mas em idiomas diferentes, sem necessariamente o auxílio de tradução de máquina (LO *et al.*, 2017). Também é necessário um dicionário léxico para cada um dos idiomas trabalhados. Dessa forma, como as sentenças entre os textos tem o mesmo valor semântico, a utilização de múltiplos dicionários léxicos aumenta a capacidade de análise. Caso um dicionário não possua a polaridade de determinado termo, ainda é possível consultar nos dicionários dos outros idiomas (ARAÚJO; PEREIRA; BENEVENUTO, 2020). A Figura 14 demonstra o fluxo de aplicação dessa abordagem.

Ela foi bastante utilizada quando os tradutores de máquina não tinham maturidade suficiente para traduzir corretamente idiomas, principalmente de origens diferentes (LO *et al.*, 2017). Como na tradução do inglês, de origem germânica, para o chinês, de origem sino-tibetana. Nesse caso há inúmeras palavras e expressões que existem em apenas um dos idiomas, e precisam ser corretamente traduzidas para não interferirem no resultado da AS. Desse modo, na abordagem baseada em corpus paralelos são utilizados dicionários léxicos de ambos os idiomas, visando resolver esse problema.

Mesmo que apresente um desempenho satisfatório na maioria dos estudos realizados, é uma abordagem que possui muitos requisitos para ser executada. A obtenção desses requisitos pode se tornar um empecilho para a aplicação dessa abordagem.

Figura 14 – Abordagem multi-idioma baseada em corpus paralelos



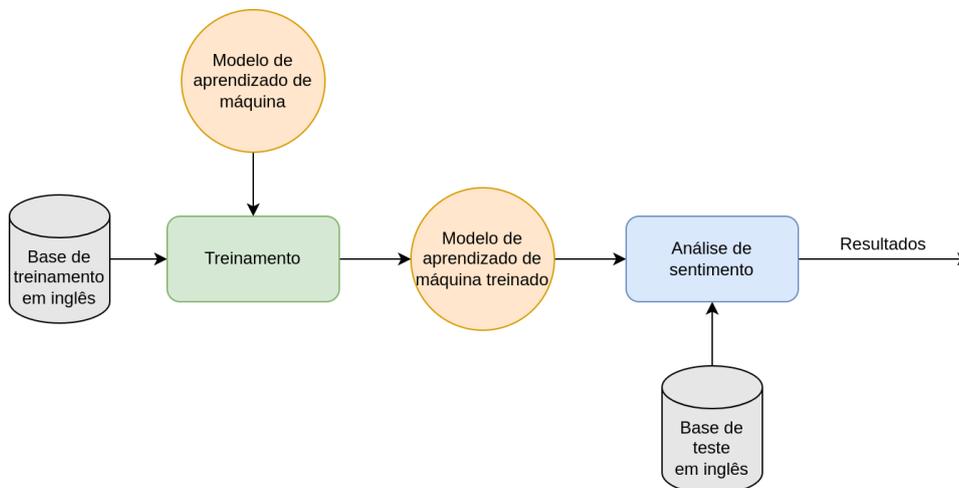
Fonte: Elaborada pelo autor, baseado em Lo *et al.* (2017).

### 3.1.3 Baseada em aprendizado de máquina

Essa abordagem utiliza-se de algoritmos de aprendizado de máquina para classificar a polaridade de um texto (ARAÚJO; PEREIRA; BENEVENUTO, 2020). Utiliza-se uma base de dados rotulada para treinar um modelo computacional a identificar características do texto que estão diretamente relacionadas à sua polaridade (Figura 15).

O modelo computacional, após treinado com uma base de dados rotulada, pode ser utilizado para identificar a polaridade de textos de outra base de dados, da qual o modelo ainda não tem conhecimento. Dessa forma, o resultado apresentado por essa abordagem está diretamente relacionado com a qualidade da base de dados utilizada para treinamento. Se o problema a ser analisado está bem definido — realizar análise de sentimento em textos — é comum utilizar-se uma base de treinamento de contexto relacionado ao problema em questão, a fim de obter-se melhores resultados.

Figura 15 – Abordagem multi-idioma baseada em aprendizado de máquina

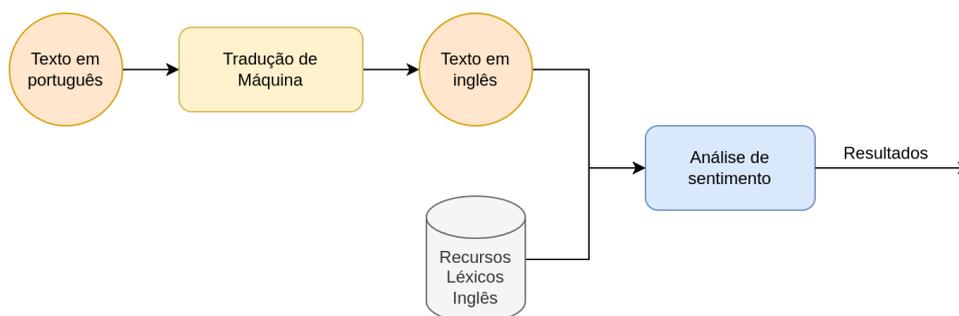


Fonte: Elaborado pelo autor (2022).

### 3.1.4 Baseada em tradução de máquina

Essa abordagem consiste em traduzir automaticamente textos, utilizando tradução de máquina, de modo a utilizar recursos léxicos de apenas um idioma (ARAÚJO; PEREIRA; BENEVENUTO, 2020). Figura 16 ilustra o fluxo de aplicação dessa abordagem.

Figura 16 – Abordagem multi-idioma baseada em tradução de máquina



Fonte: Elaborado pelo autor (2022).

Textos de idiomas com poucos recursos léxicos são traduzidos para idiomas ricos em recursos, como o inglês, por exemplo. Dessa forma é possível converter um problema multi-idiomas para um de único idioma e usufruir de melhores recursos para realizar a análise de sentimento.

Essa abordagem tem ganhado bastante espaço na análise de sentimento multi-idioma, pois apresenta bons desempenhos, facilidade na implementação e rapidez na obtenção de resultados, dado que é possível utilizar um modelo já treinado em um idioma para realizar toda a tarefa.

A atual maturidade dos tradutores de máquina possibilita uma tradução muito mais assertiva, considerando inúmeras variáveis como contexto e formalidade do texto (LO *et al.*, 2017). Essa evolução possibilitou que o uso de tradutores de máquina para a AS alcançasse desempenhos similares as demais abordagens, necessitando muito menos requisitos, tempo e trabalho.

## 3.2 TRABALHOS RELACIONADOS

De modo a verificar as abordagens de análise de sentimento multi-idiomas sendo utilizadas na literatura, alguns trabalhos foram analisados. O trabalho realizado por Sattar *et al.* (2021) implementa uma abordagem utilizando aprendizado profundo e *transformers*. O trabalho realizado por Wehrmann *et al.* (2017) implementa um modelo de *embeddings* ao nível de caracteres. O trabalho realizado por Araújo, Pereira e Benevenuto (2020) analisa a aplicação de ferramentas de tradução de máquina na análise de sentimento multi-idiomas.

### 3.2.1 *A Multi-Layer Network for Aspect-Based Cross-Lingual Sentiment Classification*

Com o objetivo de obter melhor desempenho na análise de sentimento multi-idioma baseada em aspecto, Sattar *et al.* (2021) implementaram uma abordagem utilizando aprendizado profundo e *transformers*. Em seus experimentos foi utilizada a base de dados *SemEval-2016 Task 5*, que contém avaliações de usuários/clientes em 8 idiomas e diversos domínios diferentes, como avaliações de restaurantes, hotéis e produtos (PONTIKI *et al.*, 2016). Sattar *et al.* (2021) focaram no domínio de avaliações de restaurantes, classificando-as em positivas ou negativas, sendo assim um problema de classificação binária. Os idiomas trabalhados foram inglês, francês, alemão e espanhol.

A abordagem proposta é definida em cinco etapas:

- Extração *Part of speech*: foram extraídas as *tags Part of speech* de cada avaliação da base de dados. Assim, cada palavra da avaliação recebeu um rótulo informando qual sua classificação gramatical.
- Preprocessamento: foram removidas as informações desnecessárias para a realização da análise de sentimento, como pontuação e *stop-words*. Após, foi realizada a tokenização das sentenças.
- Mapeamento entre idiomas: utilizando dicionários multi-idiomas, foi realizado o mapeamento das palavras de um idioma para os demais.
- *Word embedding*: foram criados os vetores de *tokens* utilizando BERT.
- Treinamento do modelo de aprendizado profundo.
- Predição do sentimento de uma avaliação.

O classificador utilizado foi a rede neural convolucional profunda (CNN). Foram utilizadas 3 camadas com 128 neurônios cada. A camada de saída contém 2 neurônios, positivo e negativo. A função de custo utilizada foi a entropia cruzada (*binary cross-entropy*). A abordagem proposta foi avaliada com diferentes configurações, e a que obteve melhores resultados foi utilizando  $epoch = 10$ ,  $batch\ size = 16$ , e função de ativação *tanh*.

A validação do modelo foi realizada utilizando validação cruzada *k-fold*, com  $k = 4$ . Ou seja, a base de dados foi dividida em 4 partes, e em cada validação 3 partes foram utilizadas para treinamento e 1 parte utilizada para teste. As métricas utilizadas foram precisão, *recall* e F1-Score.

Os resultados obtidos foram comparados com outras abordagens consideradas estado da arte em análise de sentimento multi-idiomas baseada em aspecto, SPC-BERT, AEN-BERT,

e LCF-BERT (Figura 17). A abordagem proposta por Sattar *et al.* (2021) obteve resultados melhores que as demais em quase todos os casos, apresentando resultado inferior que as demais em três casos:

- Testando sentenças em inglês no modelo treinado em espanhol: apresentou precisão de 58,53%, enquanto o modelo SPC-BERT apresentou 61,41%. Uma diferença relativa de 4,92%.
- Testando sentenças em alemão no modelo treinado em espanhol: apresentou precisão de 49,09%, enquanto o modelo AEN-BERT obteve 50,06%. Uma diferença relativa de 1,97%.
- Testando sentenças em inglês no modelo treinado em francês: apresentou precisão de 58,31%, enquanto o modelo LCF-BERT obteve 59,92%. Uma diferença relativa de 2,76%.

Sattar *et al.* (2021) atribuiu essa perda de desempenho aos dicionários de Espanhol-Inglês, Espanhol-Alemão e Francês-Alemão, afirmando que eles são eram contextualmente ricos.

Figura 17 – Valores de precisão, *recall* e F1-Score obtidos por método

Training Language	Testing Language	SPC-BERT			AEN-BERT			LCF-BERT			CNN-BERT		
		Pre	Rec	F1									
English	Dutch	35.53%	34.87%	35.20%	50.78%	50.52%	50.65%	54.07%	49.13%	51.48%	<b>63.67%</b>	<b>58.62%</b>	<b>61.04%</b>
	French	35.59%	35.38%	35.48%	52.11%	50.74%	51.42%	49.03%	47.95%	48.48%	<b>56.03%</b>	<b>54.65%</b>	<b>55.33%</b>
	Spanish	38.67%	34.84%	36.66%	50.61%	51.93%	51.26%	51.49%	53.98%	52.71%	<b>58.10%</b>	<b>57.72%</b>	<b>57.91%</b>
	<b>Average</b>	<b>36.60%</b>	<b>35.03%</b>	<b>35.78%</b>	<b>51.17%</b>	<b>51.06%</b>	<b>51.11%</b>	<b>51.53%</b>	<b>50.35%</b>	<b>50.89%</b>	<b>59.27%</b>	<b>57.00%</b>	<b>58.09%</b>
Spanish	English	<b>61.41%</b>	50.05%	55.15%	56.11%	49.52%	52.61%	49.84%	49.69%	49.76%	58.53%	<b>53.30%</b>	<b>55.79%</b>
	Dutch	39.31%	42.03%	40.63%	<b>50.06%</b>	46.22%	48.06%	41.26%	43.89%	42.53%	49.09%	<b>48.21%</b>	<b>48.65%</b>
	French	38.52%	41.94%	40.16%	55.91%	50.93%	53.30%	43.85%	46.46%	45.12%	<b>56.01%</b>	<b>54.34%</b>	<b>55.16%</b>
	<b>Average</b>	<b>46.41%</b>	<b>44.67%</b>	<b>45.31%</b>	<b>54.03%</b>	<b>48.89%</b>	<b>51.33%</b>	<b>44.98%</b>	<b>46.68%</b>	<b>45.81%</b>	<b>54.54%</b>	<b>51.95%</b>	<b>53.20%</b>
Dutch	English	52.96%	59.71%	56.13%	54.30%	54.90%	54.60%	48.18%	52.09%	50.06%	<b>57.24%</b>	<b>60.87%</b>	<b>59.00%</b>
	Spanish	41.50%	44.52%	42.96%	53.66%	54.21%	53.93%	44.24%	43.12%	43.67%	<b>56.02%</b>	<b>56.79%</b>	<b>56.40%</b>
	French	44.57%	44.08%	44.32%	49.38%	43.91%	46.48%	44.00%	40.79%	42.33%	<b>54.72%</b>	<b>53.88%</b>	<b>54.30%</b>
	<b>Average</b>	<b>46.34%</b>	<b>49.44%</b>	<b>47.80%</b>	<b>52.45%</b>	<b>51.01%</b>	<b>51.67%</b>	<b>45.47%</b>	<b>45.33%</b>	<b>45.36%</b>	<b>55.99%</b>	<b>57.18%</b>	<b>56.57%</b>
French	English	42.35%	48.51%	45.22%	52.43%	50.74%	51.57%	<b>59.92%</b>	54.56%	57.11%	58.31%	<b>56.04%</b>	<b>57.15%</b>
	Spanish	49.35%	50.86%	50.09%	56.75%	59.09%	57.90%	55.56%	59.36%	57.40%	<b>62.23%</b>	<b>63.72%</b>	<b>62.97%</b>
	Dutch	47.69%	43.84%	45.68%	50.55%	50.91%	50.73%	46.47%	46.66%	46.56%	<b>54.48%</b>	<b>53.18%</b>	<b>53.82%</b>
	<b>Average</b>	<b>46.46%</b>	<b>47.74%</b>	<b>47.00%</b>	<b>53.24%</b>	<b>53.58%</b>	<b>53.40%</b>	<b>53.98%</b>	<b>53.53%</b>	<b>53.69%</b>	<b>58.34%</b>	<b>57.65%</b>	<b>57.98%</b>
<b>Average</b>		<b>37.09%</b>	<b>37.65%</b>	<b>37.27%</b>	<b>43.13%</b>	<b>41.56%</b>	<b>42.29%</b>	<b>39.33%</b>	<b>39.53%</b>	<b>39.39%</b>	<b>45.92%</b>	<b>45.26%</b>	<b>45.57%</b>

Fonte: Sattar *et al.* (2021)

### 3.2.2 A character-based convolutional neural network for language-agnostic Twitter sentiment analysis

Nesse trabalho, Wehrmann *et al.* (2017) propõem um modelo de análise de sentimento agnóstico a idioma, utilizando aprendizado profundo e *embedding* ao nível de caracteres. É essa última característica que dá nome ao modelo proposto, Conv-Char-S. Diferentemente de outros métodos de análise de sentimento multi-idiomas, o método proposto por Wehrmann *et al.* (2017) não utiliza nenhuma tradução de texto. Em seu modelo, uma camada da rede neural é responsável por gerar o dicionário e extrair as características necessárias do texto.

Os experimentos foram realizados utilizando uma base de dados com 1,6 milhões de *tweets* (prevenientes da rede social Twitter), de 13 idiomas diferentes e manualmente rotulados entre as classes positiva, neutra e negativa. Os idiomas selecionados por Wehrmann *et al.* (2017) foram inglês, espanhol, português e alemão. Para adequar a base ao problema binário de análise de sentimento, os autores removeram as amostras com sentimento neutro, ficando apenas com amostras positivas e negativas. Ao final a base de dados utilizada continha 128 mil *tweets*.

O classificador utilizado foi uma rede neural CNN. Foi utilizado o algoritmo otimizador de Adam e a função de custo a log-verossimilhança negativa. As métricas avaliadas foram acurácia e F1-Score.

Os resultados obtidos pelo método proposto foram comparados com outros cinco métodos existentes na literatura: LSTM-Emb, Conv-Emb, Conv-Emb-Freeze, Conv-Char, e SVM-based. O modelo proposto por Wehrmann *et al.* (2017) obteve no geral melhor acurácia e F1-Score que os demais (Tabela 1). Porém, quando analisados os idiomas individualmente, o modelo LSTM-Emb apresentou resultados superiores para os idiomas inglês e espanhol, como pode ser observado na Tabela 2. Contudo, mesmo sendo superado nesses idiomas, os autores demonstraram que o modelo Conv-Char-S pode ser sim uma opção para qualquer desses 4 idiomas trabalhados, considerando os bons resultados obtidos e a facilidade de implementação do modelo.

Tabela 1 – Valores de acurácia e F1-Score obtidos por método.

Método	Acurácia	F1
SVM-U	0.694	0.751
SVM-UB	0.672	0.746
SVM-UBT	0.660	0.743
SVM-B	0.562	0.717
SVM-T	0.556	0.712
Conv-Emb-Freeze	0.688	0.728
Conv-Emb	0.715	0.747
LSTM-Emb	0.711	0.752
Conv-Char	0.696	0.721
Conv-Char-S [ours]	<b>0.720</b>	<b>0.756</b>

Fonte: Wehrmann *et al.* (2017)

### 3.2.3 A comparative study of machine translation for multilingual sentence-level sentiment analysis

Nesse trabalho, Araújo, Pereira e Benevenuto (2020) comparam abordagens de análise de sentimento multi-idiomas utilizando tradução de máquina com abordagens nativas, assim como analisam o desempenho das ferramentas de tradução de máquina. São analisadas 4 ferramentas de tradução de máquina: Google Translate, Microsoft Translator Text API, Yandex Translate e um programa de tradução *offline*. Os métodos de análise de sentimento analisados

Tabela 2 – Valores de acurácia e F1-Score obtidos por método e idioma.

Método	Métrica	Inglês	Alemão	Português	Espanhol
Conv-Emb-Freeze	Acurácia	0.734	0.694	0.671	0.678
	F1-Score	0.736	0.752	0.583	0.795
Conv-Emb	Acurácia	0.766	0.733	0.695	0.693
	F1-Score	0.763	0.782	0.608	0.804
LSTM-Emb	Acurácia	<b>0.770</b>	0.734	0.686	<b>0.697</b>
	F1-Score	<b>0.779</b>	0.783	0.619	<b>0.806</b>
Conv-Char	Acurácia	0.736	0.718	0.688	0.668
	F1-Score	0.729	0.758	0.589	0.776
Conv-Char-S [ours]	Acurácia	0.762	<b>0.755</b>	<b>0.706</b>	0.694
	F1-Score	0.771	<b>0.797</b>	<b>0.647</b>	0.799

Fonte: Wehrmann *et al.* (2017)

são 18, em que 15 são para o idioma inglês e 3 para idiomas específicos: Emoticons, Vader, SOCAL, Umigon, Panas-t, AFINN, Google SA, OpinionLexicon, MPQA, Emolex, Stanford, ML-Sentistrength, NRCHashtag, SASA e Happiness Index, Semantria, IBM Watson, Sentistrength. Os idiomas analisados são 14: chinês, alemão, espanhol, grego, croata, hindi, tcheco, holandês, francês, língua crioula haitiana, inglês, português, russo e italiano (ARAÚJO; PEREIRA; BENEVENUTO, 2020).

A metodologia utilizada por Araújo, Pereira e Benevenuto (2020) é composta por 3 principais etapas: identificação dos principais métodos de análise de sentimento, aquisição de bases de dados para todos os idiomas trabalhos, e aquisição valores de referências para métodos de análise de sentimento para idiomas não ingleses. Para identificar os métodos de análise de sentimento, Araújo, Pereira e Benevenuto (2020) recorreram aos autores de trabalhos relacionados com análise de sentimento em inglês. Foram listados 13 métodos, 9 deles utilizando dicionários léxicos, 2 utilizando aprendizado de máquina, e 2 de abordagem híbrida. A aquisição das bases de dados também foi realizada consultando autores de trabalhos relacionados, sendo obtidas bases dos 14 idiomas. Essas bases são compostas de sentenças rotuladas manualmente, como positiva, negativa ou neutra, conforme a polaridade do sentimento expresso. A origem dessas sentenças é diversificada, sendo algumas bases são compostas por *tweets*, e outras por textos de um assunto específico.

A partir da hipótese criada por Araújo, Pereira e Benevenuto (2020) de que “métodos de análise de sentimento no idioma inglês, quando combinados com boas ferramentas de tradução de máquina, podem obter desempenho tão bom quanto métodos nativos de outros idiomas”, alguns experimentos foram realizados de modo a validá-la. Os resultados obtidos foram analisados através das métricas de Macro-F1 e Aplicabilidade, e os experimentos foram guiados por 5 perguntas:

- Qual o impacto da ferramenta de tradução de máquina na análise de sentimento?

- Qual o desempenho de métodos para o idioma inglês quando utilizados em idiomas não ingleses traduzidos automaticamente?
- A abordagem de tradução de máquina é melhor que a abordagem nativa?
- Há diferença de desempenho nas abordagens quando avaliadas separadamente as polaridades positiva e negativa?
- Em que casos as abordagens nativas são melhores que a abordagem de tradução de máquina?

Inicialmente Araújo, Pereira e Benevenuto (2020) compararam o desempenho das ferramentas de tradução de máquina utilizadas. Todas as bases de dados foram traduzidas para o inglês utilizando as 4 ferramentas de tradução, com exceção da base original em inglês, utilizada como valor de referência. Os resultados não apresentaram diferença estatisticamente significativa quando utilizado  $\alpha = 0.95$ , sendo todos resultados iguais no intervalo de confiança (Tabela 3). Foi realizada a análise de sentimento de todas as 501 mil sentenças traduzidas, e 38 mil (7%) tiveram sua polaridade alterada durante a classificação na abordagem de tradução de máquina.

Tabela 3 – Resultado da análise de sentimento em cada ferramentas de tradução de máquina.

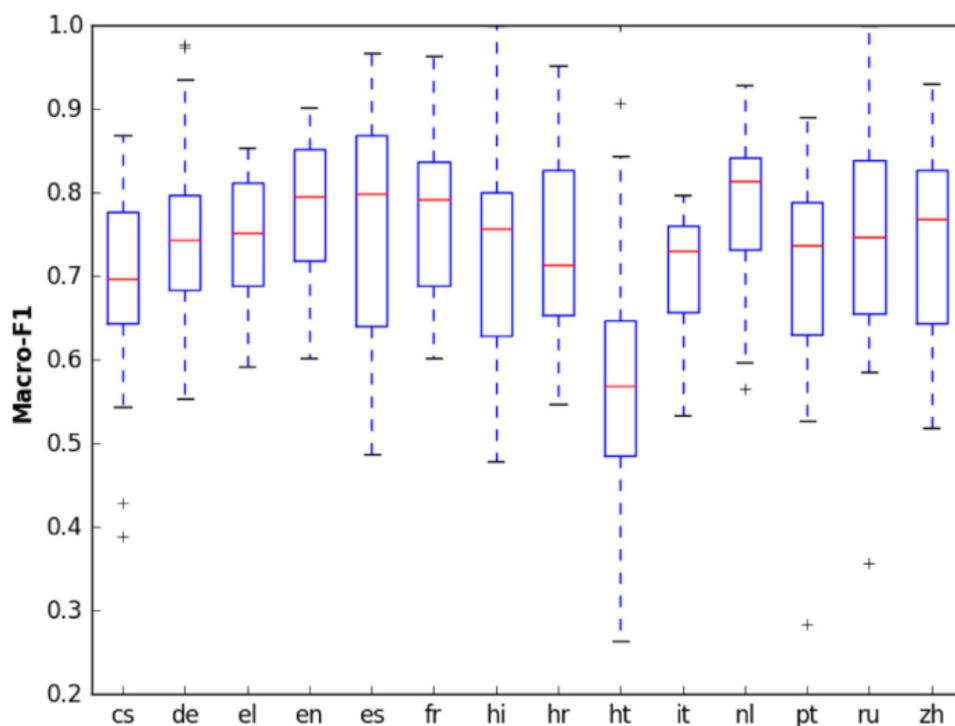
Método de Tradução	Macro-F1	Aplicabilidade
Google	0.73 $\pm$ (0.02)	0.47 $\pm$ (0.04)
Yandex	0.73 $\pm$ (0.02)	0.54 $\pm$ (0.03)
Bing	0.72 $\pm$ (0.02)	0.53 $\pm$ (0.03)
Baseline	0.70 $\pm$ (0.02)	0.51 $\pm$ (0.03)

Fonte: Araújo, Pereira e Benevenuto (2020)

Após foi aplicada a análise de sentimento por tradução de máquina em todos os 14 idiomas trabalhados. Os resultados indicaram não haver perda de eficiência na análise de desempenho quando utilizada a abordagem de tradução de máquina (Figura 18). A exceção é a língua crioula haitiana, que seu baixo desempenho é explicado por Araújo, Pereira e Benevenuto (2020) pelo fato de ser um idioma pouco utilizado fora do Haiti, dessa forma os recursos existentes para o treinamento das ferramentas de tradução é menos rico que dos demais idiomas.

Os autores concluem que as ferramentas de tradução existentes no mercado hoje são confiáveis o suficiente para não impactarem significativamente na análise de sentimento dos textos traduzidos (ARAÚJO; PEREIRA; BENEVENUTO, 2020). Os métodos de tradução de máquina utilizados para a análise de sentimento tiveram resultados próximos aos métodos nativos, sendo uma boa alternativa para casos em que seja necessário realizar a análise de sentimento em múltiplos idiomas. Os autores definem os métodos SOCAL e Sentistrength como os melhores para serem utilizado com abordagem de tradução de máquina para textos não ingleses. Também

Figura 18 – Desempenho da análise de sentimento por tradução de máquina nos idiomas trabalhados.



Fonte: Araújo, Pereira e Benevenuto (2020)

é concluído que ambas as abordagens obtiveram melhores resultados na classificação da polaridade positiva que negativa. Por fim, é implementado um sistema de análise de sentimento (*iFeel 3.0*) utilizando a linguagem de programação Java e o *framework* Spring. São implementados muitos dos métodos apresentados no trabalho, suportando análise de sentimento multi-idiomas.

## 4 SISTEMA DE ANÁLISE DE SENTIMENTO MULTI-IDIOMAS

Este capítulo apresenta os métodos propostos para o desenvolvimento de um sistema de análise de sentimento multi-idioma. A Seção 4.1 detalha a metodologia de experimentação proposta, as abordagens utilizadas e medidas de desempenho utilizadas. A Seção 4.2 descreve a base de dados utilizada. A Seção 4.3 apresenta os resultados e discussões.

### 4.1 METODOLOGIA

A metodologia utilizada neste trabalho é baseada no trabalho de Araújo, Pereira e Benevenuto (2020), em que é inicialmente levantada uma hipótese e, a partir de experimentos, pretende-se verificar sua validade.

Muitos autores já realizaram estudos com a aplicação da análise de sentimento em textos traduzidos para inglês através de tradução de máquina (ARAÚJO; PEREIRA; BENEVENUTO, 2020). Shalunts, Backfried e Commeignes (2016) aplicaram a análise de sentimento em textos traduzidos dos idiomas alemão, russo e espanhol para o inglês, por exemplo. Sua abordagem obteve resultados muito semelhantes com abordagens específicas para esses idiomas, chamadas abordagens nativas, sendo que o pior caso apresentou um desempenho inferior de 5%.

Isso indica que a utilização de tradução de máquina para aplicação da análise de sentimento no idioma inglês pode ser vantajosa, pois alinha a abundante quantidade de recursos disponíveis para esse idioma, como também o grande leque de abordagens possíveis e a possibilidade de treinar um modelo de aprendizado de máquina em um único idioma (LO *et al.*, 2017).

Pensando nisso, esse trabalho pretende validar a hipótese de que a aplicação da análise de sentimento em textos em inglês, traduzidos por tradução de máquina, possa obter resultados semelhantes às abordagens nativas para textos em outros idiomas. Então, ao final desse trabalho, deve ser possível responder as seguintes perguntas:

1. Qual o impacto do tradutor de máquina escolhido?
2. Qual é o desempenho das abordagens de análise de sentimento em textos de idiomas diferentes do inglês quando aplicada a tradução de máquina?
3. Em que casos a abordagem de tradução de máquina apresenta melhor resultado que as abordagens nativas de análise de sentimento?
4. Há diferença de desempenho quando realizada a análise de sentimento, separadamente, em avaliações de polaridade positiva e negativa?

O sistema de análise de sentimento proposto para a execução dos experimentos utiliza um modelo BERT para cada um dos idiomas analisados: inglês, alemão, espanhol e francês. Esses modelos já são pré-treinados em seus respectivos idiomas, porém nenhum deles é destinado especificamente para a tarefa de análise de sentimento. Para obter-se os modelos aplicáveis na análise de sentimento, foi necessário realizar o *fine-tuning*.

Os modelos BERT utilizados e suas principais características podem ser visualizados na Tabela 4. Todos os modelos utilizados foram baseados no BERT-base<sup>1</sup>, por isso foram implementados com 12 camadas principais (*layers*), 768 camadas ocultas (*hidden layers*) e 12 núcleos de atenção (*attention heads*).

Tabela 4 – Modelo BERT utilizado em cada idioma.

Idioma	Modelo	Tamanho do Dicionário	Qtd. de Hiperparâmetros	Qtd. de dados de treinamento
Inglês	Base BERT Cased	29k	110M	25 GB
Alemão	German BERT Cased	30k	110M	12 GB
Espanhol	BETO Cased	31k	110M	-
Francês	Camembert Base Cased	32k	110M	138 GB

Fonte: Elaborado pelo autor (2022).

### 4.1.1 Tokenização

A primeira etapa de processamento do BERT é a tokenização dos textos de entrada. Nesse processo foram utilizados tokenizadores que fazem a diferenciação entre letras maiúsculas e minúsculas, *cased-tokenizers*. Isso porque é relevante para a análise de sentimento identificar termos em maiúsculo digitados pelo usuário, pois eles tendem a intensificar a polaridade passada pela avaliação. Como a utilização do termo “HORRÍVEL” para enfatizar a má experiência com algum produto ou em casos positivos o termo “ÓTIMO”, por exemplo. A tokenização foi realizada utilizando o dicionário léxico do modelo de cada idioma (Tabela 4).

Os modelos BERT utilizados suportam até 512 *tokens* nos *embeddings* de entrada. Contudo, para não exceder o limite computacional, foi necessário limitar o número de *tokens*, encontrando um número que abrangesse a maioria dos dados. Para isso, foram gerados os *tokens* de todas as avaliações da base de teste do idioma inglês (Figura 19a). Aplicando-se a operação logarítmica nos dados, é possível perceber os dados são distribuídos normalmente (Figura 19b). Visto que os dados possuem uma distribuição log-normal, estatísticas descritivas como a média e desvio padrão foram estimadas. Verificou-se que a média de *tokens* é 44,5 *tokens* por avaliação, com um desvio padrão de 42,15 *tokens*. Visando incluir a maioria dos *tokens* das avaliações, um limiar foi selecionado com base no desvio padrão em relação à média (BUSSAB;

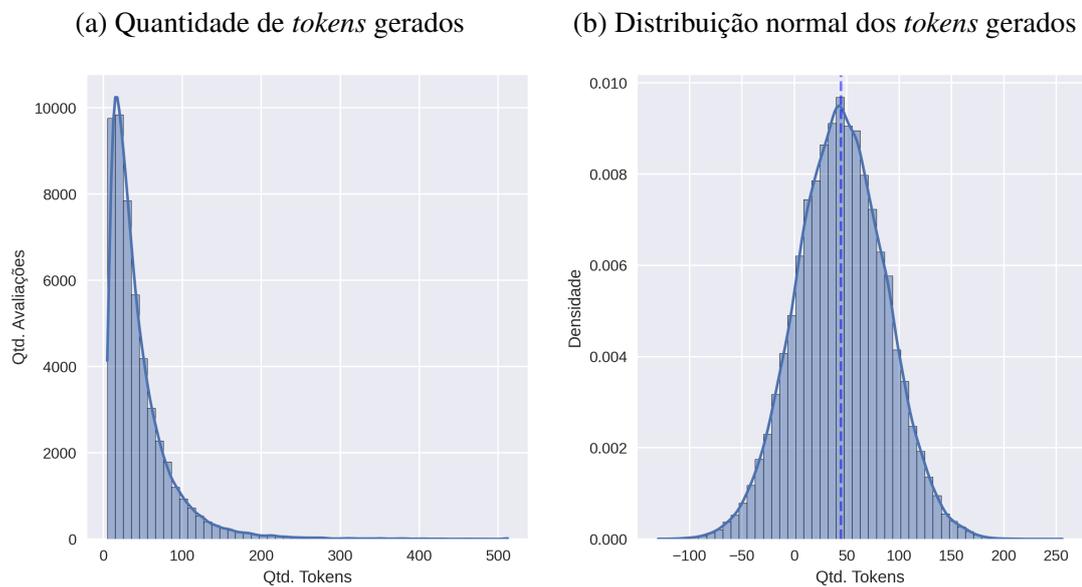
<sup>1</sup> Modelo BERT treinado com menor número de camadas e parâmetros quando comparado ao BERT-large (DEVLIN *et al.*, 2018).

MORETTIN, 2004). Considerando que a limitação de *tokens* incluía até 97,5% deles nas avaliações, o limiar foi selecionado com o valor da média mais duas vezes o desvio padrão (Tabela 5). O *tokenizador* foi configurado para limitar as avaliações em até 130 *tokens*<sup>2</sup>.

Tabela 5 – Estatísticas dos *tokens* gerados

Média $\mu$	44.50
Desvio Padrão $\sigma$	42.15
Limiar ( $\mu + 2 * \sigma$ )	128.81
Fonte: Elaborado pelo autor (2022).	

Figura 19 – Quantidade de *tokens* gerados e sua distribuição



Fonte: Elaborado pelo autor (2022).

#### 4.1.2 Obtenção do modelo para a análise de sentimento

Os modelos a serem utilizados na análise de sentimento foram obtidos através do *fine-tuning* dos modelos BERT bases de cada idioma. Foi realizado o *fine-tuning* no modelo BERT dos idiomas trabalhados, utilizando a base de treinamento do respectivo idioma. O *fine-tuning* foi realizado através de um programa implementado na linguagem de programação Python, através da ferramenta Pytorch e bibliotecas como Numpy, Pandas e Sklearn. A execução foi realizada em uma máquina com processador AMD Ryzen 5 3500x, 16 GB de RAM e placa de vídeo NVIDIA GTX 1650 com 4GB de memória. O *fine-tuning* de cada modelo, utilizando 50 mil registros, levou cerca de cinco horas.

O processo de *fine-tuning* de cada modelo foi executado 10 vezes (10 *epochs*), utilizando uma taxa de aprendizado (*learning rate*) de  $2e - 4$ . Nos experimentos realizados, essa

<sup>2</sup> Utilizado o parâmetro *max\_length=130* no tokenizador.

configuração obteve melhores resultados que a configuração utilizada por Devlin *et al.* (2018), de  $2e - 5$  para taxa de aprendizado e 3 *epochs*. O algoritmo de otimização utilizado foi o *Stochastic Gradient Descent* (SGD), pois necessita de menos recursos da máquina e apresentou resultados muito próximos do algoritmo Adam, utilizado por Devlin *et al.* (2018), que leva consideravelmente mais tempo para convergir e extrapolou o limite computacional.

Também foi necessário definir o método de validação dos modelos gerados. Nesse ponto foram avaliados os métodos de validação cruzada e *Stratified Hold Out* no modelo BERT em inglês. Inicialmente foi utilizado o método de validação cruzada *K-Fold*, com o valor de  $k = 5$ . Então a base de treinamento foi dividida em 5 partes: 4 partes foram utilizadas para treinar os modelos, e 1 parte para avaliar. Após realizado o *fine-tuning* dos modelos, eles foram avaliados na base de dados de teste, e o resultado do melhor modelo foi armazenado para posterior comparação. Então, foi realizado o *fine-tuning* do modelo utilizando o método *Stratified Hold Out*, em que toda a base de treinamento é utilizada. O modelo foi avaliado na mesma base de teste utilizando o método de validação cruzada. Então os resultados obtidos foram comparados (Tabela 6). O modelo utilizando *Stratified Hold Out* obteve resultado superior ao do melhor modelo gerado com validação cruzada. Sendo assim, decidiu-se continuar os experimentos utilizando *Stratified Hold Out*.

Tabela 6 – Comparação dos métodos de validação

	Validação Cruzada	<i>Stratified Hold Out</i>
<i>Acc.</i>	84%	85.55%
<i>Prec.</i>	80%	85.46%
<i>Rec.</i>	88%	85,75%
<i>F1.</i>	84%	85.55%

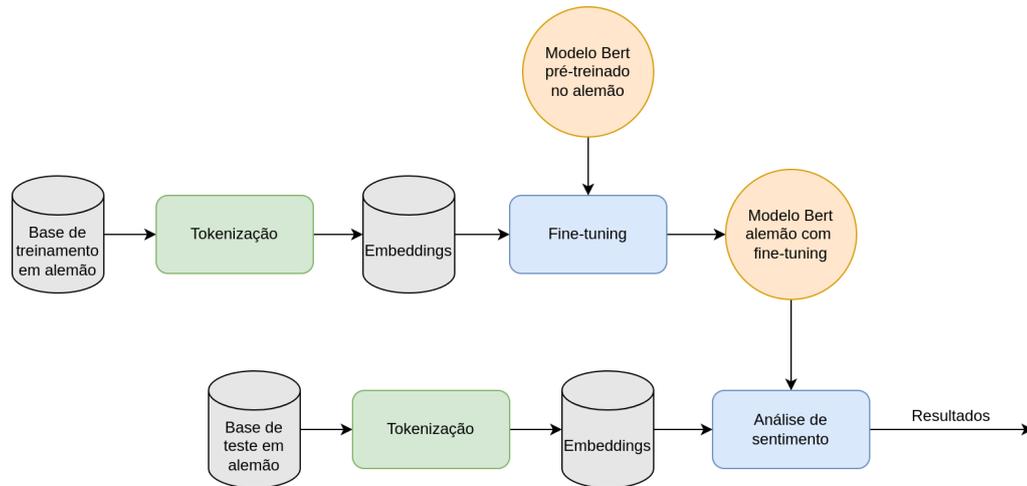
Fonte: Elaborado pelo autor (2022).

### 4.1.3 Abordagens utilizadas

A metodologia proposta aplica a análise de sentimento nos experimentos em duas abordagens distintas: nativa e tradução de máquina.

A abordagem nativa tem por objetivo realizar a análise de sentimento nos diferentes modelos BERT de cada idioma (Tabela 4). É uma abordagem baseada em aprendizado de máquina (Seção 3.1.3), em que cada modelo é treinado e validado com dados do seu respectivo idioma. O diagrama da abordagem nativa é apresentado na Figura 20, exemplificado no idioma alemão. Inicialmente é realizada a tokenização da base de treinamento do idioma a ser trabalhado. Assim, são gerados os vetores de *embeddings* e, com o modelo BERT pré-treinado para o idioma, é realizado o *fine-tuning* do modelo. Após essa etapa, o modelo está pronto para ser utilizado na tarefa de análise de sentimento. Dessa forma, o modelo é avaliado na base de teste do respectivo idioma, que também passa pelo processo de tokenização. Gerados os *embeddings* da base de teste, e de posse do modelo BERT já com *fine-tuning*, é realizada a análise de sentimento.

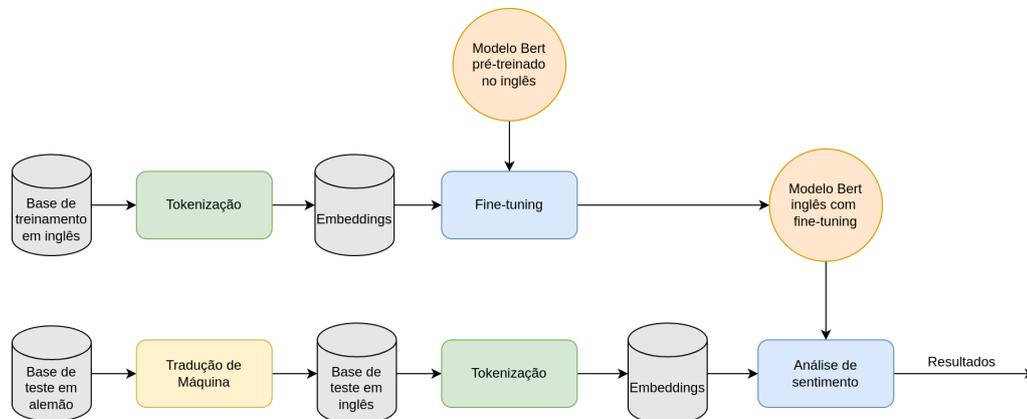
Figura 20 – Diagrama da abordagem nativa



Fonte: Elaborado pelo autor (2022).

A abordagem de tradução de máquina tem por objetivo realizar a análise de sentimento utilizando apenas o modelo BERT do idioma inglês. É uma abordagem baseada em tradução de máquina, onde o modelo é treinado e validado sempre no idioma inglês. A Figura 21 apresenta o diagrama da abordagem de tradução de máquina, exemplificado no idioma alemão. Inicialmente é realizada a tokenização da base de treinamento do idioma inglês. Assim, são gerados os vetores de *embeddings* e, com o modelo BERT pré-treinado para o idioma inglês, é realizado o *fine-tuning* do modelo. Dado que essa abordagem visa utilizar apenas um modelo BERT, de idioma inglês, a base de teste dos demais idiomas precisa ser traduzida. A tradução das bases de teste foram realizadas através das ferramentas de tradução de máquina disponíveis. De posse das bases de teste traduzidas para o inglês, é realizado o processo de tokenização. Gerados os *embeddings* da base de teste, e de posse do modelo BERT já com *fine-tuning*, é realizada a análise de sentimento.

Figura 21 – Diagrama da abordagem de tradução de máquina



Fonte: Elaborado pelo autor (2022).

## 4.2 BASE DE DADOS

As abordagens foram avaliadas utilizando a base de dados *Multilingual Amazon Reviews Corpus* (MARC), a qual possui avaliações de usuários sobre produtos comprados na plataforma da Amazon. Essas avaliações foram coletadas dos *marketplaces* da Amazon nos países Estados Unidos, Japão, Alemanha, França, Espanha e China, no respectivo idioma de cada país (KEUNG *et al.*, 2020).

As avaliações coletadas estão diretamente relacionadas a uma compra efetuada na plataforma, enviadas durante o período de 1º de novembro de 2015 a 1º de novembro de 2019 (KEUNG *et al.*, 2020). Visando evitar criar um viés na base de dados, a base foi construída selecionando no máximo 20 avaliações por produto e do mesmo usuário. Produtos com apenas uma avaliação e avaliações com menos de 20 caracteres também não foram consideradas na base de dados.

Pelo fato de avaliações em inglês serem frequentemente encontradas em todos os *marketplaces* da Amazon, Keung *et al.* (2020, p2) aplicaram um algoritmo de detecção de idioma em todas as avaliações coletadas. Assim, foi possível filtrar as avaliações para a base de dados do país conter apenas avaliações do seu respectivo idioma. A base de dados em inglês contém apenas avaliações do *marketplace* dos Estados Unidos escritas em inglês, a base de dados em alemão contém apenas avaliações do *marketplace* da Alemanha escritas em alemão, e assim com todos os idiomas trabalhados.

As avaliações também foram filtradas pelo vocabulário utilizado pelo usuário, excluídas avaliações que contenham termos que não estejam presentes em pelo menos 20 outras avaliações do mesmo idioma (KEUNG *et al.*, 2020). Considerando que as avaliações na plataforma da Amazon são divididas numa escala de 5 estrelas, foi realizada uma nova seleção das avaliações da base de dados de tal forma que cada estrela componha 20% das amostras. Por fim, todas as avaliações foram limitadas a 2.000 caracteres.

Além da avaliação, diversas informações sobre a avaliação (meta-informações) é disponibilizada pela base. A estrutura de cada avaliação na base de dados é mostrada no Figura 22. Das informações presentes em cada registro, foram utilizadas nesse trabalho a pontuação (*stars*) e o texto da avaliação (*review\_body*).

A base de dados MARC possui um total de 1.260.000 avaliações. Cada idioma possui 210.000 avaliações. Para padronizar as avaliações de sistemas de análise de texto, cada idioma é dividido em 3 grupos: 200.000 avaliações para treinamento, 5.000 para desenvolvimento e 5.000 para teste (KEUNG *et al.*, 2020). A Tabela 7 apresenta algumas estatísticas das avaliações por idioma: número de produtos, número de usuários e média de caracteres por avaliação da base de dados de cada idioma. O número de produtos e usuários considerados em cada idioma é bem próximo, indicando que as avaliações não ficaram concentradas em determinado produto ou usuário (KEUNG *et al.*, 2020).

Figura 22 – Estrutura de cada avaliação na base de dados

```
{
  "review_id": "en_0000258",
  "reviewer_id": "reviewer_en_0010355",
  "product_id": "product_en_0000097",
  "language": "en",
  "stars": 5,
  "review_title": "Salad Spinner",
  "review_body": "Perfect for herbs and leafy vegetables!",
  "product_category": "kitchen"
}
```

Fonte: Keung *et al.* (2020)

Tabela 7 – Estatísticas da base de treinamento.

	<b>Ing.</b>	<b>Ale.</b>	<b>Esp.</b>	<b>Fra.</b>
Número de produtos	196.745	189.148	179.076	183.345
Número de usuários	185.541	171.620	150.938	157.922
Média de caracteres por avaliação	178	207	151	159

Fonte: Adaptado de Keung *et al.* (2020).

Como o problema de análise de sentimento é binário, a escala de pontuação de 5 estrelas foi utilizada para gerar a polaridade do sentimento (i.e., negativa e positiva). Para isso, as avaliações com pontuação 1 e 2 estrelas foram consideradas negativas. Avaliações com pontuação 4 e 5 estrelas foram consideradas positivas. As avaliações com pontuação 3 estrelas foram consideradas neutras e assim removidas da base de dados. Dessa forma, a base de teste resultante é composta por 4.000 avaliações. Como a demanda computacional é grande para o treinamento de modelos baseados em redes neurais profundas (HAYKIN, 2007), a base de treinamento foi reduzida a 50.000 avaliações (25.000 para cada polaridade).

De modo a executar os experimentos propostos nesse trabalho, as bases de testes dos idiomas alemão, espanhol e francês, foram traduzidas para o inglês, utilizando as ferramentas de tradução do Google (*Cloud Translation API*) e Microsoft (*Microsoft Translator*). Dessa forma, cada um desses idiomas possui três bases: a original, em inglês traduzida pelo Google e em inglês traduzida pela Microsoft.

Todos tratamentos na base de dados foram realizados por *scripts* implementados na linguagem Python. Para a tradução da base de dados foram utilizadas as APIs disponibilizadas por cada ferramenta de tradução.

## 4.3 RESULTADOS E DISCUSSÕES

Inicialmente, aplicando a abordagem nativa, foi construído um modelo para cada idioma utilizando os dados do próprio idioma. Esses modelos foram estimados através da realização do *fine-tuning* no modelo BERT do respectivo idioma. Nesta operação foram utilizadas 50.000 avaliações, divididas igualmente entre as duas polaridades. Para avaliar o modelo estimado, foram utilizadas 4.000 avaliações, divididas igualmente entre as duas polaridades. A Tabela 8 mostra o desempenho dos modelos obtidos para cada idioma aplicando a abordagem nativa.

Tabela 8 – Desempenho da abordagem nativa em cada idioma.

Idioma original	<i>Acc</i>	<i>Prec.</i>	<i>Rec.</i>	<i>F1</i>	<i>MCC</i>
Inglês	89,17%	84,74%	95,55%	89,13%	0.790
Alemão	90,70%	88,57%	93,45%	90,69%	0.815
Espanhol	91,50%	87,96%	96,15%	91,48%	0.834
Francês	88,87%	96,92%	80,30%	88,79%	0.789

Fonte: Elaborado pelo autor (2022).

De posse dos resultados da abordagem nativa em cada idioma, os experimentos das seções a seguir visam responder as questões propostas na metodologia.

### 4.3.1 Avaliação dos tradutores de máquina

Esta seção apresenta os resultados do sistema de análise de sentimentos proposto a partir de entradas traduzidas por duas ferramentas de tradução de máquina — Google ou Microsoft — para diferentes idiomas. Todos os idiomas foram avaliados com as diferentes traduções das bases de dados para o idioma inglês no sistema desenvolvido com base na abordagem nativa do idioma inglês. Os resultados dos experimentos usando a base do idioma inglês e as bases traduzidas por diferentes ferramentas de tradução são mostradas na Tabela 9.

Os resultados mostram não haver uma diferença estatisticamente significativa entre as ferramentas de tradução. As diferenças de acurácia foram avaliadas com teste Z (BUSSAB; MORETTIN, 2004) para diferença entre proporções com um nível de significância de 5% (isto é,  $\alpha = 0.05$ ). Como não há diferença, as bases traduzidas pela ferramenta da Microsoft foram utilizadas nos demais experimentos.

Os resultados obtidos para os idiomas traduzidos foram próximos ao resultado do idioma inglês. A diferença entre o resultado do idioma alemão ( $Acc = 87,99\%$ ) com o idioma inglês não é estatisticamente significativa, enquanto a diferença entre o idioma espanhol ( $Acc = 86,80\%$ ) e o inglês foi significativa. Uma explicação para isso é a origem dos idiomas. Os idiomas alemão e inglês compartilham a mesma origem (germânica) (FULK, 2008) e tendem a produzir resultados semelhantes, enquanto o idioma espanhol, de origem latina, não teria tais

Tabela 9 – Desempenho da análise de sentimentos para as traduções de diferentes idiomas para o inglês. Os resultados para a língua inglesa são mostrados como referência.

Idioma original	Ferramenta de tradução	<i>Acc</i>	<i>Prec.</i>	<i>Rec.</i>	<i>F1</i>	<i>MCC</i>
Inglês		89,17%	84,74%	95,55%	89,13%	0.790
Alemão	Google	87,15%	84,30%	91,30%	87,12%	0.746
	Microsoft	87,99%	83,35%	94,94%	87,93%	0.767
Espanhol	Google	86,92%	82,77%	93,25%	86,87%	0.744
	Microsoft	86,80%	81,45%	95,30%	86,70%	0.747
Francês	Google	92,37%	92,35%	92,40%	92,37%	0.847
	Microsoft	93,17%	92,55%	93,90%	93,17%	0.864

Fonte: Elaborado pelo autor (2022).

semelhanças (CLACKSON, 2011), por isso teria pior resultado em um modelo do idioma inglês. Entretanto, o idioma francês alcançou resultados melhores que o do próprio idioma inglês e isso pode derrubar a alegação anterior. Assim, é necessário investigar o motivo desta diferença.

Outra característica analisada nas bases traduzidas é o número de *tokens* gerados por avaliação (Tabela 10). Como a base traduzida do idioma francês possui em média menos *tokens* por avaliação, pode-se concluir que o limite de 130 *tokens* por avaliação não remove *tokens* das avaliações. Pode-se ver na Tabela 10 que para o idioma francês, mais de 99% da distribuição ( $\mu + 3 * \sigma$ ) encontra-se abaixo de 117 *tokens*. Esta não remoção dos *tokens* pode estar contribuindo para o melhor desempenho apresentado. Apesar disso, o mesmo resultado não é obtido para o idioma espanhol. Pode-se especular que o resultado está relacionado com o idioma espanhol possuir uma remoção maior que a do francês (média e desvio padrão maiores que as do francês) e a sua origem diferente da língua inglesa.

Tabela 10 – Estatísticas dos *tokens* por avaliação das bases traduzidas para o inglês.

Idioma original	Média ( $\mu$ )	Desvio Padrão ( $\sigma$ )	$\mu + 2 * \sigma$	$\mu + 3 * \sigma$
Alemão	44.45	37.15	118.74	155.89
Espanhol	37.48	31.16	99.81	130.97
Francês	34.14	27.36	88.86	116.22

Fonte: Elaborado pelo autor (2022).

### 4.3.2 Comparação dos resultados da abordagem nativa e de tradução de máquina

Essa seção apresenta os resultados obtidos da análise de sentimento utilizando as abordagens nativas e de tradução de máquina propostas. Os resultados são apresentados na Tabela 11.

Tabela 11 – Desempenho das abordagens nativas e tradução de máquina

Idioma original	Abordagem	<i>Acc</i>	<i>Prec.</i>	<i>Rec.</i>	<i>F1</i>	<i>MCC</i>
Alemão	Nativa	90,70%	88,57%	93,45%	90,69%	0.815
	Tradução	87,99%	83,35%	94,94%	87,93%	0.767
Espanhol	Nativa	91,50%	87,96%	96,15%	91,48%	0.834
	Tradução	86,80%	81,45%	95,30%	86,70%	0.747
Francês	Nativa	88,87%	96,92%	80,30%	88,79%	0.789
	Tradução	93,17%	92,55%	93,90%	93,17%	0.864

Fonte: Elaborado pelo autor (2022).

Os idiomas alemão e espanhol apresentam resultados superiores na abordagem nativa. A diferença do resultado da acurácia (*Acc*) entre as abordagens foi avaliada com teste Z (BUSSAB; MORETTIN, 2004), e, para um nível de significância de 5%, é estatisticamente significativa. A maior diferença entre as abordagens foi observada no idioma espanhol, onde a tradução obteve um resultado relativo aproximadamente 5% inferior à abordagem nativa (91,5%). Ao analisar a métrica de *recall*, a diferença apresentada entre as abordagens não é significativa. Contudo, a métrica de precisão é significativamente superior nas abordagens nativas, apresentando a maior diferença no modelo espanhol, 6,51%. Esse comportamento das métricas de precisão e *recall* sugere que ambas abordagens possuem taxas de acerto próximas para a polaridade positiva, porém na polaridade negativa as abordagens nativas possuem taxas de acerto melhores. A análise sobre a assertividade entre as polaridades é apresentada na Seção 4.3.3. Uma explicação para a diferença de resultado entre as abordagens pode ser o tamanho do dicionário léxico utilizados por esses modelos, maiores que o utilizado no modelo inglês. Outra explicação é a quantidade de dados de utilizados no pré-treinamento, que no caso do modelo espanhol foi muito superior aos demais modelos (Tabela 4).

O idioma francês apresentou pior resultado na abordagem nativa. Esse desempenho foi impactado pelo baixo *recall* ( $Rec = 80,30\%$ ). Isso indica que o modelo para o francês não conseguiu classificar corretamente 19,70% das avaliações de polaridade positiva. Contudo, a abordagem de tradução de máquina no idioma francês apresenta *recall* próximo aos demais idiomas ( $Rec = 93,90\%$ ). Isso indica que o *recall* baixo da abordagem nativa não está diretamente relacionado com a base de teste utilizada. Dessa forma, é preciso analisar o modelo base utilizado para o idioma francês (Camembert Base Cased) para verificar se alguma configuração do próprio modelo está gerando um viés de classificação.

### 4.3.3 Impacto da polaridade no desempenho

Essa seção apresenta os resultados obtidos da análise de sentimento utilizando as abordagens nativas e de tradução de máquina quando avaliadas as polaridades separadamente. Os resultados são apresentados na Tabela 12. Esse experimento avalia bases de dados que contém apenas 1 polaridade, dessa forma, as métricas utilizadas em problemas de classificação binária não são aplicadas (Seção 2.6). Então, é utilizada a taxa de acerto, que expressa a porcentagem de acertos do classificador para a polaridade avaliada.

Tabela 12 – Desempenho das abordagens quando avaliadas as polaridades separadamente.

Idioma	Abordagem	Negativa	Positiva
Inglês	Nativa	82,80%	95,55%
Alemão	Nativa	87,95%	93,45%
	Tradução	81,05%	94,94%
Espanhol	Nativa	86,85%	96,15%
	Tradução	78,30%	95,30%
Francês	Nativa	97,44%	80,30%
	Tradução	92,44%	93,90%

Fonte: Elaborado pelo autor (2022).

Os idiomas inglês, alemão e espanhol apresentam, em ambas abordagens, taxa de acerto superior na classificação das avaliações de polaridade positiva. Na abordagem de tradução de máquina, o idioma francês apresenta mesmo comportamento que os demais, obtendo melhores resultados nas avaliações positivas. Contudo, o francês na abordagem nativa apresentou melhores resultados na classificação de avaliações negativas. Isso colabora com a hipótese de o modelo base utilizado nesse idioma (Camembert Base Cased) possuir algum viés de polaridade.

O texto de algumas avaliações erroneamente classificadas foram analisados manualmente. Percebeu-se que muitas das avaliações de polaridade negativa apresentam sentimentos de ambas as polaridades, utilizando-se muito de conjunções adversativas como “mas”, “porém” e “contudo” para realizar a troca de polaridade de sentimento no texto. Essas avaliações começam expressando um sentimento positivo sobre o produto, porém finalizam com um sentimento negativo. O usuário, talvez por conta da sua frustração, marcou a avaliação como negativa na plataforma de compra. Para exemplificar essa característica do texto, foram selecionadas 2 avaliações de polaridade negativa erroneamente classificadas, retiradas da base de dados do idioma espanhol e alemão, respectivamente, e traduzidas para o português pelo Google Tradutor:

- “O case é bom, protege o ipad. A única desvantagem é que chegou com a tela trincada e a caixa aberta.”
- “Fiquei muito feliz com o relógio até que a pulseira caiu após 6 meses de uso regular.”

Nas análises realizadas nas avaliações de polaridade positiva erroneamente classificadas, percebeu-se que muitas delas apresentam sentimento positivo sobre o produto em geral, mas sentimento negativo sobre alguma característica específica do produto ou experiência da compra, como a entrega, por exemplo. Essas avaliações utilizam termos negativos como “infelizmente”, “raiva” e apresentam o uso constante de “não”. Para exemplificar isso, foram selecionadas 2 avaliações positivas erroneamente classificadas, retiradas da base de dados do idioma espanhol e alemão, respectivamente, e traduzidas para o português pelo Google Tradutor:

- “Isso foi perfeito, mas me deixa com raiva porque eles geralmente não são cuidadosos com os livros, com outros já aconteceu comigo que a capa estava dobrada.”
- “O CD é muito bom, mas infelizmente tem um pequeno arranhão. Como o CD é para presente, estou um pouco decepcionado.”

## 5 CONSIDERAÇÕES FINAIS

O avanço do uso de plataformas digitais para consumo de conteúdo ou realizações de compras está mudando a forma como a sociedade expressa sua opinião. Hoje, grande parte das opiniões expressas sobre determinado produto ou serviço pode ser encontrada nessas plataformas, através de avaliações de usuários. Visando adquirir informações sobre essas avaliações, tarefas de NLP podem ser aplicadas, como a análise de sentimento.

De modo a solucionar problemas existentes nas tarefas de NLP, novos algoritmos, arquiteturas e modelos têm surgido. Arquiteturas como o *transformer* e modelos como o BERT surgiram para otimizar os resultados de tarefas de NLP, e hoje são consideradas estado da arte nessa área. A análise de sentimento se beneficiou disso e vem aplicando modelos BERT para obter melhores resultados, como na aplicação da análise de sentimento em idiomas diferentes do inglês. Contudo, o modelo BERT foi originalmente proposto para o idioma inglês, sendo posteriormente implementado para outros idiomas por autores diferentes.

Nesse contexto, este trabalho apresenta uma comparação dos resultados da aplicação da análise de sentimento multi-idiomas, utilizando os modelos BERT implementados para os idiomas inglês, alemão, espanhol e francês, e aplicando as abordagens de análise de sentimento nativas e de tradução de máquina. Para isso, foi implementado um sistema de análise de sentimento utilizando os modelos BERT e realizados os experimentos propostos. Inicialmente verificou-se o impacto da ferramenta de tradução de máquina na análise de sentimento utilizando a abordagem de tradução. Outro experimento avaliou o desempenho da análise de sentimento utilizando as duas abordagens propostas: nativa e de tradução de máquina. A base de dados MARC foi utilizada para avaliar o desempenho do sistema.

Os resultados demonstraram não haver diferenças significativas entre as traduções produzidas por ambas as ferramentas. Assim, a escolha da ferramenta de tradução de máquina não tem impacto no resultado da análise de sentimento.

Os experimentos mostraram que a abordagem nativa apresentou melhor desempenho que a abordagem de tradução de máquina. A maior diferença ocorreu no idioma espanhol, em que a abordagem de tradução de máquina teve um desempenho relativo aproximadamente 5% inferior. Contudo, a abordagem nativa necessita de treinamento e *fine-tuning* de um modelo por idioma a ser trabalhado, tornando sua implementação mais complexa e demandando mais poder computacional. Dessa forma, dependendo dos requisitos da aplicação do sistema, a diferença de desempenho entre as abordagens pode não ser significativa a ponto de justificar o uso da abordagem nativa. Principalmente se os idiomas trabalhados forem de mesma origem linguística, onde a abordagem de tradução de máquina tende a obter melhores resultados, como no caso do alemão e inglês. A aplicação da abordagem de tradução de máquina também pode ser vantajosa

em idiomas que não possuem bons recursos léxicos disponíveis para a aplicação da abordagem nativa. Ao analisar o desempenho das abordagens quando testadas separadamente, a polaridade positiva foi melhor classificada nos idiomas inglês, alemão e espanhol, demonstrando que ainda há pontos a serem aprimorados em NLP para tratar ambas polaridades igualmente.

Como trabalhos futuros, pode-se analisar os resultados do modelo BERT francês (Cambert base cased) para entender a origem da diferença dos resultados obtidos, quando comparados com os demais modelos. Também pretende-se avaliar o modelo BERT multi-idíomas (BERT Multilingual Cased), treinado em mais de 100 idiomas.

## REFERÊNCIAS

- ABBASI, A.; CHEN, H.; SALEM, A. Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. **ACM Trans. Inf. Syst.**, Association for Computing Machinery, New York, NY, USA, v. 26, n. 3, jun 2008. ISSN 1046-8188. Disponível em: <<https://doi.org/10.1145/1361684.1361685>>.
- AGARWAL, B.; MITTAL, N. Optimal feature selection for sentiment analysis. In: GELBUKH, A. (Ed.). **Computational Linguistics and Intelligent Text Processing**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. p. 13–24. ISBN 978-3-642-37256-8.
- ARAÚJO, M.; PEREIRA, A.; BENEVENUTO, F. A comparative study of machine translation for multilingual sentence-level sentiment analysis. **Information Sciences**, v. 512, p. 1078–1102, 2020. ISSN 0020-0255. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0020025519309879>>.
- ASGHAR, D. M. *et al.* A review of feature extraction in sentiment analysis. **Journal of Basic and Applied Research International**, v. 4, p. 181–186, 01 2014.
- BAHDANAU, D.; CHO, K.; BENGIO, Y. **Neural Machine Translation by Jointly Learning to Align and Translate**. arXiv, 2014. Disponível em: <<https://arxiv.org/abs/1409.0473>>.
- BIRJALI, M.; KASRI, M.; BENI-HSSANE, A. A comprehensive survey on sentiment analysis: Approaches, challenges and trends. **Knowledge-Based Systems**, v. 226, p. 107134, 2021. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S095070512100397X>>.
- BUSSAB, W.; MORETTIN, P. **Estatística básica**. Saraiva, 2004. (Metodos quantitativos para economistas e administradores). Disponível em: <<https://books.google.com.br/books?id=ARMJQgAACAAJ>>.
- CAMBRIA, E. Affective computing and sentiment analysis. **IEEE Intelligent Systems**, v. 31, n. 2, p. 102–107, 2016.
- CHICCO, D.; JURMAN, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. **BMC Genomics**, v. 21, n. 1, p. 6, 2020. ISSN 1471-2164. Disponível em: <<https://doi.org/10.1186/s12864-019-6413-7>>.
- CLACKSON, J. Introduction. In: \_\_\_\_\_. **A Companion to the Latin Language**. [S.l.]: John Wiley & Sons, Ltd, 2011. cap. 1, p. 1–6. ISBN 9781444343397.
- DASHTIPOUR, K. *et al.* Multilingual sentiment analysis: State of the art and independent comparison of techniques. **Cognitive Computation**, v. 8, p. 757–771, 2016.
- DEVLIN, J. *et al.* **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**. arXiv, 2018. Disponível em: <<https://arxiv.org/abs/1810.04805>>.
- FELDMAN, R. Techniques and applications for sentiment analysis. **Commun. ACM**, v. 56, p. 82–89, 04 2013.
- FULK, R. D. English as a germanic language. **Companion to the history of the English language**, p. 142–49, 2008.

GANGADHARAN, V. *et al.* Paraphrase detection using deep neural network based word embedding techniques. In: **2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184)**. [S.l.: s.n.], 2020. p. 517–521.

GHANNAY, S. *et al.* Word embedding evaluation and combination. In: **Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)**. Portorož, Slovenia: European Language Resources Association (ELRA), 2016. p. 300–305. Disponível em: <<https://aclanthology.org/L16-1046>>.

HAYKIN, S. **Redes Neurais: Princípios e Prática**. [S.l.]: Artmed, 2007. ISBN 9788577800865.

KEUNG, P. *et al.* The multilingual amazon reviews corpus. In: **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing**. [S.l.: s.n.], 2020.

LIU, B. Sentiment analysis and opinion mining. **Synthesis Lectures on Human Language Technologies**, v. 5, n. 1, p. 1–167, 2012. Disponível em: <[doi:10.2200/S00416ED1V01Y201204HLT016](https://doi.org/10.2200/S00416ED1V01Y201204HLT016)>.

LO, S. L. *et al.* Multilingual sentiment analysis: from formal to informal and scarce resource languages. **Artificial Intelligence Review**, v. 48, 12 2017.

MADHOUSHI, Z.; HAMDAN, A. R.; ZAINUDIN, S. Sentiment analysis techniques in recent works. In: **2015 Science and Information Conference (SAI)**. [S.l.: s.n.], 2015. p. 288–291.

MEDHAT, W.; HASSAN, A.; KORASHY, H. Sentiment analysis algorithms and applications: A survey. **Ain Shams Engineering Journal**, v. 5, n. 4, p. 1093–1113, 2014. ISSN 2090-4479. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S2090447914000550>>.

MISHEV, K. *et al.* Evaluation of sentiment analysis in finance: From lexicons to transformers. **IEEE Access**, v. 8, p. 131662–131682, 2020.

NASSIRTOUSSI, A. K. *et al.* Text mining for market prediction: A systematic review. **Expert Systems with Applications**, v. 41, n. 16, p. 7653–7670, 2014. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0957417414003455>>.

OLIVAS, E. S. *et al.* **Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques**. 1. ed. Information Science Reference, 2009. ISBN 1605667668; 9781605667669. Disponível em: <[libgen.li/file.php?md5=4e8ba85af7deb52584bacbdb07d85f7](http://libgen.li/file.php?md5=4e8ba85af7deb52584bacbdb07d85f7)>.

PONTIKI, M. *et al.* SemEval-2016 task 5: Aspect based sentiment analysis. In: **Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)**. San Diego, California: Association for Computational Linguistics, 2016. p. 19–30. Disponível em: <<https://aclanthology.org/S16-1002>>.

SATTAR, K. *et al.* A multi-layer network for aspect-based cross-lingual sentiment classification. **IEEE Access**, v. 9, p. 133961–133973, 2021.

SHALUNTS, G.; BACKFRIED, G.; COMMEIGNES, N. The impact of machine translation on sentiment analysis. In: . [S.l.: s.n.], 2016.

SHI, Y. *et al.* Intrinsic or extrinsic evaluation: An overview of word embedding evaluation. In: **2018 IEEE International Conference on Data Mining Workshops (ICDMW)**. [S.l.: s.n.], 2018. p. 1255–1262.

SHOBHA, G.; RANGASWAMY, S. Chapter 8 - machine learning. In: GUDIVADA, V. N.; RAO, C. (Ed.). **Computational Analysis and Understanding of Natural Languages: Principles, Methods and Applications**. Elsevier, 2018, (Handbook of Statistics, v. 38). p. 197–228. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0169716118300191>>.

TABOADA, M. Sentiment analysis: An overview from linguistics. **Annual Review of Linguistics**, v. 2, n. 1, p. 325–347, 2016.

THAKKAR, A.; CHAUDHARI, K. A comprehensive survey on deep neural networks for stock market: The need, challenges, and future directions. **Expert Systems with Applications**, v. 177, p. 114800, 2021. ISSN 0957-4174. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0957417421002414>>.

UPRETI, B. *et al.* Knowledge-driven approaches for financial news analytics. In: \_\_\_\_\_. [S.l.: s.n.], 2019. p. 375–404. ISBN 978-981-13-8318-2.

VASWANI, A. *et al.* Attention is all you need. In: GUYON, I. *et al.* (Ed.). **Advances in Neural Information Processing Systems**. Curran Associates, Inc., 2017. v. 30. Disponível em: <<https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf>>.

WEHRMANN, J. *et al.* A character-based convolutional neural network for language-agnostic twitter sentiment analysis. In: **2017 International Joint Conference on Neural Networks (IJCNN)**. [S.l.: s.n.], 2017. p. 2384–2391.

WIEBE, J. *et al.* Learning subjective language. **Comput. Linguist.**, MIT Press, Cambridge, MA, USA, v. 30, n. 3, p. 277–308, sep 2004. ISSN 0891-2017. Disponível em: <<https://doi.org/10.1162/0891201041850885>>.