

Análise de Sentimentos: Impacto da Tradução Neural na Avaliação de Desempenho

Lincoln Wallace Valentim da Costa Silva



CENTRO DE INFORMÁTICA
UNIVERSIDADE FEDERAL DA PARAÍBA

João Pessoa, PB

2023

Lincoln Wallace Valentim da Costa Silva

Análise de Sentimentos:
Impacto da Tradução Neural na Avaliação de Desempenho

Monografia apresentada ao curso
Ciência de dados e Inteligência Artificial
do Centro de Informática, da Universidade Federal da Paraíba,
como requisito para a obtenção do grau de Bacharel

Orientador: Yuri de Almeida Malheiros Barbosa

Dezembro de 2023

Catálogo na publicação
Seção de Catalogação e Classificação

S586a Silva, Lincoln Wallace Valentim da Costa.

Análise de sentimentos: impacto da tradução neural
na avaliação de desempenho / Lincoln Wallace Valentim
da Costa Silva. - João Pessoa, 2023.
36 f. : il.

Orientação: Yuri Malheiros.
TCC (Graduação) - UFPB/CI.

1. Processamento de linguagem natural. 2. Tradução
neural. 3. Análise de sentimentos. 4. Métricas de
avaliação. 5. Classificador de sentimentos. I.
Malheiros, Yuri. II. Título.

UFPB/CI

CDU 004



CENTRO DE INFORMÁTICA
UNIVERSIDADE FEDERAL DA PARAÍBA

Trabalho de Conclusão de Curso de Ciência de Dados e Inteligência Artificial intitulado ***Análise de Sentimentos: Impacto da Tradução Neural na Avaliação de Desempenho*** de autoria de Lincoln Wallace Valentim da Costa Silva, aprovada pela banca examinadora constituída pelos seguintes professores:

Prof. Dr. Yuri de Almeida Malheiros Barbosa
Universidade Federal da Paraíba (UFPB)

Prof. Dr. Thaís Gaudencio do Rêgo
Universidade Federal da Paraíba (UFPB)

Prof. Dr. Tiago Maritan Ugulino de Araújo
Universidade Federal da Paraíba (UFPB)

João Pessoa, 11 de dezembro de 2023

Centro de Informática, Universidade Federal da Paraíba
Rua dos Escoteiros, Mangabeira VII, João Pessoa, Paraíba, Brasil CEP: 58058-600
Fone: +55 (83) 3216 7093 / Fax: +55 (83) 3216 7117

Eudaimonia: é como se chama a sensação de ser tomado por um sentimento bom sem explicação. É aquela vontade de viver que surge quando menos se espera e mais se precisa. É o universo nos dando um abraço de esperança. É sentir que somos o melhor de nós. É aquela sensação de ano-novo de que tudo pode ser melhor, e é.

AGRADECIMENTOS

One Piece é uma série de mangá que narra a história de um jovem em busca do maior tesouro pirata que existe. Em sua jornada, ele enfrenta diversas dificuldades, experiências frustrantes, momentos de alegria e muitas amizades preciosas. Ao longo dessa aventura, ele cruza o caminho de pessoas que se tornam seus amigos, com os quais compartilha essa incrível experiência. Entretanto, acredito que no final dessa série ele vai descobrir que o verdadeiro “One Piece” são os amigos que fizemos ao longo dessa jornada.

Gostaria de iniciar meus agradecimentos pela minha família, que sempre esteve ao meu lado ao longo desses anos. Minha mãe, que conhece todos os meus altos e baixos, meu irmão, que, mesmo distante, nunca deixou de me apoiar, e meu pai. Além disso, quero expressar minha gratidão aos amigos que se tornaram minha segunda família, o NA Data Science. Nossa jornada foi marcada por risadas, desafios, separações, mas nunca deixamos ninguém para trás. Não poderia deixar de agradecer as outras amizades que surgiram ao longo do caminho e acreditaram sempre no meu potencial de conseguir terminar, mesmo sem nunca ter visto algumas pessoalmente e visto outras poucas vezes, sempre me animaram com palavras.

Um agradecimento especial aos meus orientadores, Yuri Malheiros e Thaís Gaudencio, pela paciência que tiveram comigo nos últimos meses. Eles sempre encontraram maneiras de me acalmar e me guiar durante os momentos mais difíceis. Encaro essa fase da minha vida como meu próprio “One Piece”, que descobri ao longo desses seis longos anos de jornada e que se encerra aqui.

RESUMO

Com a crescente facilidade de acesso à internet nos últimos anos, houve um aumento significativo na geração de conteúdo pelos usuários, incluindo avaliações de produtos e análises em redes sociais, entre outros. Esse cenário rico em dados se mostrou benéfico para a análise de sentimentos. No entanto, muitos idiomas enfrentam uma carência de recursos e conjuntos de dados para conduzir análises de sentimentos eficazes. Para lidar com essas lacunas de dados, por exemplo na língua portuguesa, este trabalho propôs o uso de tradução neural automática como uma estratégia para obter dados de análise de sentimentos de idiomas mais amplamente explorados, como o inglês. O estudo apresentou dois cenários experimentais para análise de sentimentos após a tradução automática de textos, um com *tweets* e outro com avaliações de produtos (*reviews*). Os resultados obtidos a partir dos *tweets* no idioma original apresentaram um F1-Score de 59,08%, uma precisão de 60,94% e uma cobertura de 59,31%, para 3 classes de sentimentos: Positivo, Negativo e Neutro. Após a tradução neural, os resultados foram um F1-Score de 53,29%, cobertura de 54,73% e precisão de 52,62%, para as mesmas classes. No cenário das avaliações de produtos (*reviews*), foram consideradas apenas duas classes de sentimentos: Negativo e Positivo, e os resultados no idioma original foram um F1-Score de 83,00%, com precisão e cobertura também em 83,00%. Após a tradução, os resultados foram um F1-Score de 80,05%, com precisão e cobertura também em 80,05%. Nesses resultados podemos observar erros no processo de tradução neural, que impossibilitaram a classificação das frases e no processo de classificação incorreta, observamos erros principalmente em frases que não apresentavam palavras relacionadas diretamente ao sentimento, sendo ele representado por toda a frase. Dessa forma, o estudo sugere que a tradução neural pode ser uma ferramenta valiosa para coletar recursos de análise de sentimentos em idiomas com recursos limitados, como uma forma de superar a falta de dados para a pesquisa em idiomas específicos.

Palavras-chave: <Processamento de Linguagem Natural>, <Tradução Neural>, <Análise de Sentimentos>, <Métricas de Avaliação>, <Classificador de Sentimentos>.

ABSTRACT

With the increasing ease of internet access in recent years, there has been a significant increase in user-generated content, including product reviews and social media analyses, among other types of content. This data-rich environment has proven to be beneficial for sentiment analysis. However, many languages face a lack of resources and datasets for effective sentiment analysis. To address these data gaps, for example in the Portuguese language, this study proposed the use of neural machine translation as a strategy to obtain sentiment analysis data from more widely explored languages, such as English. The study presented two experimental scenarios for sentiment analysis after automatic neural translation of texts, one with tweets and the other with product reviews. The results obtained from the original tweets in the Portuguese language showed an F1-Score of 59.08%, precision of 60.94%, and recall of 59.31% for three sentiment classes: Positive, Negative, and Neutral. After neural translation, the results showed an F1-Score of 53.29%, recall of 54.73%, and precision of 52.62% for the same classes. In the scenario of product reviews, where only two sentiment classes, Negative and Positive, were considered, the results in the original language yielded an F1-Score of 83.00%, with precision and recall also at 83.00%. After translation, the results showed an F1-Score of 80.05%, with precision and recall also at 80.05%. These results indicate errors in the neural translation process that hindered sentence classification and incorrect classification, particularly in sentences that did not contain words directly related to sentiment, as sentiment was represented by the entire sentence. Thus, the study suggests that neural translation can be a valuable tool for gathering sentiment analysis resources in languages with limited resources, as a means to overcome the lack of data for research in specific languages.

Key-words: <Natural Language Processing>, <Neural Translation>, <Sentiment Analysis>, <Assessment Metrics>, <Feelings Classifier>.

LISTA DE FIGURAS

1	Arquitetura do Modelo M2M-100.	17
2	Exemplo de Matriz de Confusão.	18
3	Fluxograma do Experimento	23
4	Matrizes de Confusão do Conjunto de Dados TweetSentBr.	29
5	Matrizes de Confusão do Conjunto de Dados B2W-Review01.	30

LISTA DE TABELAS

1	Exemplos de Análise de Sentimentos.	16
2	Trabalhos Relacionados.	22
3	Exemplos de <i>Tweets</i> do Conjunto de Dados TweetSentBR.	24
4	Exemplos de <i>Reviews</i> do Conjunto de Dados B2W.	24
5	Resultados das Métricas de Avaliação do Conjunto de Dados TweetSentBr.	27
6	Resultados das Métricas de Avaliação do Conjunto de Dados B2W-Review01.	28
7	Exemplos de Classificações Neutras Rotuladas Positivas(Falsos Positivos).	29
8	Exemplos de Classificações Neutras Rotuladas Negativas(Falsos Positivos).	30
9	Exemplos de Classificações Neutras Rotuladas Positivas(Falsos Positivos).	31
10	Exemplos de Classificações Neutras Rotuladas Negativas(Falsos Positivos).	31
11	Erros de Tradução.	32

LISTA DE ABREVIATURAS

- EUA - Estados Unidos da América
- FN - Falso Negativo
- FP - Falso Positivo
- PLN - Processamento de Linguagem Natural
- RNN - Redes Neurais Recorrentes
- TA - Tradução Neural ou Tradução automática
- TN - Verdadeiro Negativo
- TP - Verdadeiro Positivo

SUMÁRIO

1	INTRODUÇÃO	13
1.1	Objetivo geral	14
1.2	Objetivos específicos	14
1.3	Estrutura da monografia	14
2	CONCEITOS GERAIS E REVISÃO DA LITERATURA	15
2.1	Processamento de Linguagem Natural	15
2.2	Análise de Sentimentos	15
2.3	Tradução Neural	16
2.4	Métricas de Avaliação	18
2.5	Trabalhos Relacionados	19
3	METODOLOGIA	23
3.1	Coleta das Bases de Dados	23
3.1.1	TweetSentBr	23
3.1.2	B2W-Review01	24
3.2	Seleção de Ferramenta de Tradução Automática	24
3.3	Tradução dos Textos	25
3.4	Classificações de Sentimentos	25
3.5	Experimentos	25
3.6	Avaliação de desempenho	26
4	APRESENTAÇÃO E ANÁLISE DOS RESULTADOS	27
5	CONCLUSÕES E TRABALHOS FUTUROS	33
	REFERÊNCIA	34

1 INTRODUÇÃO

A facilidade de acesso à Internet nos últimos anos resultou em um aumento significativo na geração de conteúdo pelos usuários, que abrange desde postagens em redes sociais até avaliações de produtos, entre outros. Esses tipos de dados possuem um valor considerável, uma vez que têm diversas aplicações, como, por exemplo, análises de opiniões sobre filmes. Conforme destacado por Rocha, Fernandes e Aguiar (2019), a plataforma *Twitter* se tornou um aliado importante na promoção e disseminação dessas análises. Outro exemplo relevante são as opiniões políticas, como demonstrado em um estudo realizado em 2018 para analisar os sentimentos dos usuários do *Twitter* em relação às eleições presidenciais, conforme mencionado por Suter et al. (2019). Esses exemplos de aplicações têm impulsionado o crescimento do campo de análise de sentimentos, uma área da inteligência artificial que busca desenvolver métodos computacionais e algoritmos capazes de analisar e classificar os sentimentos presentes em textos, palavras e frases.

Contudo, muitos desses conteúdos podem conter informações irrelevantes para os usuários. Para lidar com esse tipo de conteúdo, entra em cena o Processamento de Linguagem Natural (PLN), uma área de pesquisa que se concentra em capacitar os computadores a manipular e interpretar textos por meio de técnicas que buscam se aproximar da linguagem humana. Em outras palavras, o PLN busca tornar os computadores capazes de compreender e interagir com a linguagem escrita de forma semelhante à compreensão humana (LIDDY, 2001).

De acordo com Hirschberg e Manning (2015), os avanços significativos nos modelos de PLN foram impulsionados por competições, que, inicialmente, eram financiadas e organizadas pelo Departamento de Defesa dos EUA, mas que, atualmente, são amplamente desenvolvidas pela comunidade de desenvolvedores. Essas competições têm desempenhado um papel fundamental no avanço da pesquisa em PLN, impulsionando o desenvolvimento de técnicas e algoritmos cada vez mais eficazes.

Um desafio constante e que deve ser levado em consideração ainda hoje, é a limitação dos idiomas, uma vez que a maioria dos recursos está disponível apenas em inglês. Apesar do português ser um dos principais idiomas falados na web, os recursos disponíveis para ele são limitados, como destacado por Pereira (2021). No entanto, o trabalho de França e Oliveira (2014) resultou na criação de um conjunto de dados em português com base nas postagens dos usuários do *Twitter* em 2013, especialmente relacionadas aos protestos, e esses dados são valiosos para análises de sentimentos.

Uma solução para esses casos de escassez de recursos é a utilização de modelos de tradução neural. Estes modelos, baseados em redes neurais, têm se mostrado bastante eficazes, como demonstrado pelo avanço tecnológico nessa área (BAHDANAU; CHO; BENGIO, 2014).

1.1 Objetivo geral

Este trabalho tem como objetivo avaliar a viabilidade do uso de tradutores neurais para permitir a utilização de modelos de análise de sentimentos em inglês na classificação de textos em português. Para alcançar esse objetivo, será empregado o modelo de tradução neural M2M-100, juntamente com o modelo *Twitter-XLM-roBERTa-base* para a classificação de sentimentos.

1.2 Objetivos específicos

Abaixo serão listados os objetivos específicos para resolução do trabalho:

- Traduzir os conjuntos de dados em português para o inglês, usando o modelo de tradução neural: M2M-100.
- Classificar os dados em português e em inglês, usando o modelo de classificação multilíngue: *Twitter-XLM-roBERTa-base*.
- Avaliar e comparar os resultados antes da tradução e após a tradução dos dados.

1.3 Estrutura da monografia

A estrutura deste trabalho foi dividida em cinco capítulos: introdução, conceitos gerais e revisão de literatura, metodologia, apresentação e análise dos resultados, conclusões e trabalhos futuros. Na introdução, apresentamos a definição do problema, o objetivo geral do estudo e os objetivos específicos. Em conceitos gerais e revisão de literatura, abordamos os conceitos importantes para o entendimento da pesquisa. Na metodologia, descrevemos os passos a serem seguidos para reproduzir os experimentos. Em seguida, na seção de apresentação e análise dos resultados, mostramos os resultados obtidos descritos na metodologia. Por fim, na seção de conclusões e trabalhos futuros, apresentamos as observações a partir dos resultados alcançados, juntamente com sugestões para trabalhos futuros.

2 CONCEITOS GERAIS E REVISÃO DA LITERATURA

Neste capítulo são apresentados conceitos gerais para o entendimento do trabalho proposto, sendo eles, Processamento de Linguagem Natural, Análise de Sentimentos, Tradução Neural, Métricas de Avaliação e Trabalhos relacionados nas suas seções 2.1, 2.2, 2.3, 2.4, 2.5, respectivamente.

2.1 Processamento de Linguagem Natural

Liddy (2001) definiu Processamento de Linguagem Natural (*PLN*) como uma área de estudo composta por técnicas computacionais voltadas para a análise e representação de textos em diversos gêneros, modos e idiomas. Seu principal objetivo é capacitar os sistemas computacionais a compreender e gerar texto de maneira semelhante à linguagem humana. Isso possibilita a execução de uma ampla variedade de tarefas de forma mais próxima ao modo como as pessoas interagem com a linguagem escrita. Essas tarefas podem incluir análise de sentimentos, tradução automática, entre outras.

2.2 Análise de Sentimentos

De acordo com Benevenuto, Ribeiro e Araújo (2015), existem duas abordagens distintas para lidar com a análise de sentimentos. A primeira abordagem envolve o uso de aprendizado de máquina, na qual um modelo de classificação é treinado com base em conjuntos de dados previamente classificados e, posteriormente, usado para classificar novos textos com base nesse treinamento. Contudo, é importante salientar que esse tipo de classificador requer uma grande quantidade de dados para obter um desempenho adequado. A segunda abordagem, por outro lado, não depende de modelos alimentados por dados, mas sim de técnicas que analisam a polaridade do texto para determinar o sentimento associado a ele. Isso significa que, em vez de usar um modelo pré-treinado, essa abordagem avalia o texto com base em indicadores de polaridade, como palavras ou expressões associadas a sentimentos positivos ou negativos.

Segundo Vinodhini e Chandrasekaran (2012), análise de sentimentos também é conhecida como mineração de opinião, este campo de estudo se dedica à análise das opiniões, avaliações e graus de satisfação em relação a produtos, serviços, organizações, eventos e outros tópicos. Ela permite a identificação de diversos sentimentos, como tristeza, felicidade ou raiva, entre outros. Porém, o principal foco desse campo reside nas opiniões que expressam ou implicam sentimentos positivos ou negativos. Diversos desafios existem nessa área de estudo, entre eles, o fato de que uma mesma frase pode expressar um sentimento distinto, dependendo do contexto que ela é empregada.

No trabalho de Brum e Nunes (2017), é proposta a criação de um conjunto de dados rotulados com três possíveis classificações: positivas, negativas e neutras. Essas classificações foram determinadas com base em um conjunto de características. Frases que expressam sensações de alegria, contentamento ou enaltecimento foram classificadas como positivas, enquanto frases que demonstram raiva, oposição ou discordância foram classificadas como negativas. Por outro lado, o rótulo neutro foi atribuído a frases em que não foi possível identificar claramente nenhum tipo de sentimento das outras duas classes.

A Tabela 1 apresenta alguns exemplos dessas classificações para uma melhor compreensão.

Texto	Classe
Por mais #Encontro com as Irmãs Galvão, adorei elas.	positiva
Tchau Victor B. É obvio que vc vai sair. #MasterChefBR	neutra
Qualquer um que saísse no time vermelho eu ficava sussa, mas do time azul, tô bad! #MasterChefBR	negativa

Tabela 1: Exemplos de Análise de Sentimentos.

Visto que a análise multilíngue vem se tornando cada vez mais atrativa, é interessante destacar o trabalho de Barbieri, Anke e Camacho-Collados (2021), que propôs um classificador baseado em Rede Neural. Este classificador foi treinado com um conjunto de 198 milhões de exemplos de *tweets* em vários idiomas. Os resultados alcançados por este classificador têm se mostrado muito satisfatórios, especialmente quando se trata da classificação de *tweets* em oito idiomas específicos que foram adequadamente ajustados.

Entretanto, observa-se uma lacuna quando se trata de análise de sentimentos em alguns idiomas, uma vez que a maioria dos métodos e recursos está disponível principalmente para o idioma inglês. Nesse contexto, o trabalho de Reis et al. (2015), propôs um processo que faz uso de um tradutor de fácil acesso, o Google Translate¹, para a tradução dos conjuntos de dados para o inglês. Posteriormente, foram usados métodos de análise de sentimentos para investigar os benefícios e malefícios da análise de sentimento multilíngue.

2.3 Tradução Neural

Segundo Wang et al. (2021) a tradução neural, também conhecida como tradução automática (TA), é um campo de estudo que busca capacitar os computadores a traduzir sentenças de um idioma para outro, independentemente dos idiomas envolvidos. De acordo com Martin (2023), esse tipo de tradução tem contribuído significativamente para o acesso à informação, permitindo a tradução de diversos tipos de conteúdo em poucas etapas, visto que grande quantidade são apresentados em inglês.

¹<https://translate.google.com.br/>

No trabalho de Dong et al. (2015) é proposto um tradutor neural capaz de lidar com múltiplos idiomas. Entretanto, devido à diversidade linguística, um tradutor que abrange muitos idiomas pode encontrar desafios para manter a eficácia de tradução, uma vez que cada idioma possui suas particularidades. Apesar dessas considerações, os avanços contínuos na pesquisa têm proporcionado resultados favoráveis na área de tradução neural. Um exemplo de tradutor neural é o modelo M2M-100, proposto no trabalho Fan et al. (2021). Esse modelo foi treinado com 100 idiomas diferentes e é baseado na arquitetura *encoder-decoder*, na Figura 1, podemos visualizar essa arquitetura do modelo.

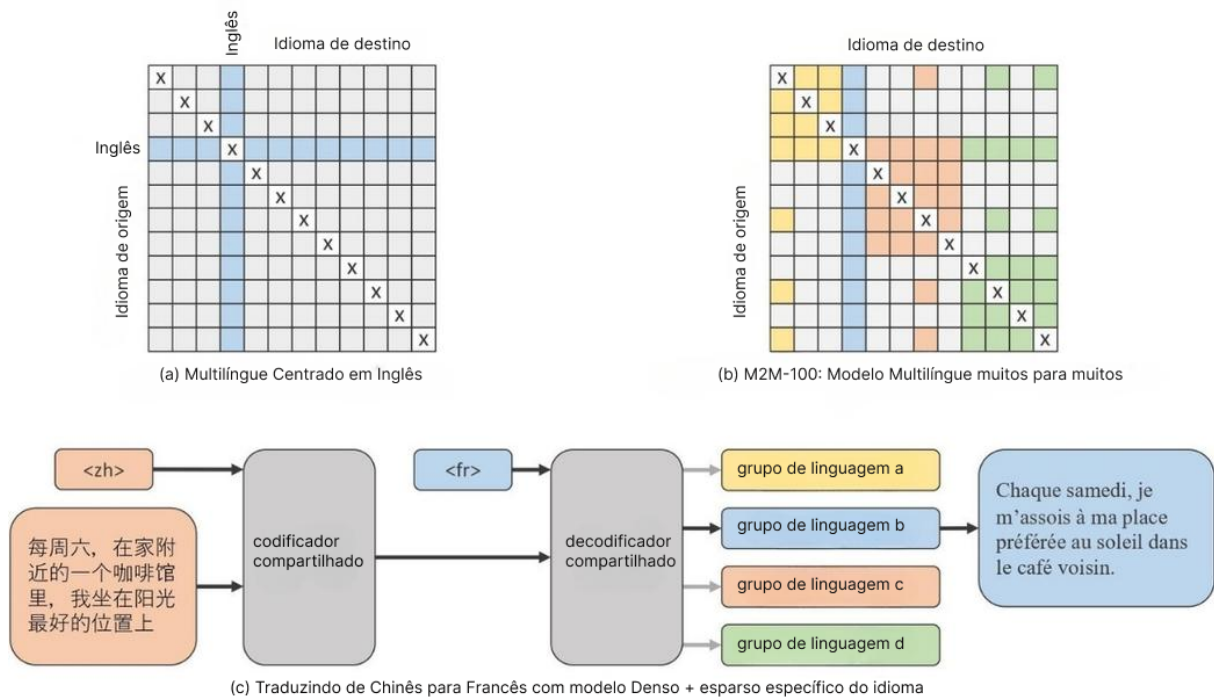


Figura 1: Arquitetura do Modelo M2M-100.

Fonte: Adaptado de (FAN et al., 2021)

Segundo o estudo de Tan et al. (2020), o *encoder* desempenha a função de organizar as entradas de texto em representações contínuas e ocultas. No contexto de modelos de tradução baseados em Redes Neurais Recorrentes (*RNN*), a rede é empregada para capturar a complexidade das frases e a ordem das palavras na sentença de origem. O estado final é utilizado para representar o texto de origem de maneira mais completa. Por outro lado, o *decoder* é responsável pela tradução do texto proveniente do *encoder*. Ele gera palavras sequencialmente, uma de cada vez, com o objetivo de criar uma sequência adequada que corresponda à frase de origem.

Conforme descrito no estudo Chatzikoumi (2020), a avaliação de um tradutor neural pode-se fazer através de métricas automatizadas, como, por exemplo, a Precisão (P) e a (C). Essa avaliação envolve a comparação dos dados de saída com os dados de entrada para determinar se a tradução é aceitável, semelhante ao processo de classificação

de sentimentos.

2.4 Métricas de Avaliação

De acordo com Reis et al. (2015), para avaliar o desempenho de um classificador de sentimentos, são normalmente utilizadas métricas de avaliação, como Precisão, Cobertura e *F1-Score*. Essas métricas são geradas a partir da matriz de confusão. Lunardi, Viterbo e Bernardini (2016) definiram essa matriz para um problema com n classes como uma matriz $n \times n$, representada por M_{ij} . Os elementos da diagonal principal ($i=j$) representam o número de acertos das classes, enquanto os elementos fora da diagonal principal ($i \neq j$) representam os valores de erros de classificação. Na Figura 2, podemos ver um exemplo dessa matriz.

		Valor Predito	
		Sim	Não
Real	Sim	Verdadeiro Positivo (TP)	Falso Negativo (FN)
	Não	Falso Positivo (FP)	Verdadeiro Negativo (TN)

Figura 2: Exemplo de Matriz de Confusão.

Fonte: Autoria Própria

Hossin e Sulaiman (2015) definiram as seguintes equações para os cálculos dessas métricas. A Precisão (P) é utilizada para medir a taxa de acertos nos valores positivos que realmente pertenciam à classe positiva. Assim, ela pode ser calculada através Equação 1:

$$P = \frac{TP}{TP + FP} \quad (1)$$

A Cobertura (C) é usada para medir a quantidade de casos positivos que foram classificados corretamente. Ela pode ser calculada através da Equação 2:

$$C = \frac{TP}{TP + TN} \quad (2)$$

O *F1-Score* é a media harmônica entre a Precisão e a Cobertura, podendo ser calculada através da Equação 3:

$$F1 - Score = \frac{2 * (P * C)}{P + C} \quad (3)$$

O valor do *F1-Score* pode variar entre 0 e 1, onde 1 indica um desempenho perfeito,

e valores menores indicam um desempenho inferior.

2.5 Trabalhos Relacionados

Ao desenvolver um classificador de sentimentos, é amplamente reconhecido que a construção de um sistema robusto demanda uma quantidade substancial e diversa de dados. Contudo, em certos idiomas, deparamo-nos com um desafio, uma vez que a disponibilidade de dados é insuficiente para esse propósito. Diante dessa situação, a tradução direta não é uma opção viável. Assim, Poncelas et al. (2020) propõe a utilização do modelo OpenNMT² de tradução neural como solução, permitindo a conversão do *feedback* dos clientes para diferentes idiomas. A presente pesquisa se dedica a investigar os possíveis efeitos da tradução de frases sobre um classificador automático de sentimentos, com o objetivo de discernir os impactos positivos ou negativos inerentes a essa abordagem. Para avaliar o desempenho do classificador, utilizaram a métrica de acurácia como uma medida avaliativa. Chegaram à conclusão de que a tradução pode, de fato, resultar em uma redução no desempenho do classificador. Isso indica que o uso de um classificador automático pode ser prejudicado pela qualidade da tradução.

Neste estudo conduzido por Shalunts, Backfried e Commeignes (2016), a investigação da influência da tradução automática na análise de sentimentos é apresentada. O estudo se baseia no uso de duas ferramentas específicas: o classificador de sentimentos multilíngue “*SentiSAIL*” e o tradutor automático “*SDL Language Weaver*”. A pesquisa se concentra na análise de notícias gerais em três idiomas distintos: russo, espanhol e alemão. Inicialmente, o artigo explora como o “*SentiSAIL*” se adapta ao comportamento de um novo idioma introduzido no estudo, que é o espanhol. Em seguida, é conduzido um experimento no qual a tradução automática é empregada para traduzir os três idiomas mencionados para o inglês. Além disso, é realizada a análise de sentimentos nos idiomas antes e depois da tradução. Os resultados deste experimento indicam resultados satisfatórios para as métricas de *Aggregation*, *Averaging* e *Maximization*, demonstrando a eficácia da abordagem que permite a utilização do classificador existente, evitando a necessidade de desenvolver modelos específicos para cada idioma.

No estudo realizado por Barriere e Balahur (2020), o foco está no aprimoramento da análise de sentimentos em *tweets* escritos em idiomas diferentes do inglês. Para alcançar esse objetivo, eles empregaram abordagens como tradução automática e o uso de classificadores tanto monolíngues quanto multilíngues. O objetivo é expandir o conjunto de dados disponíveis para análise, superando o desafio da limitada disponibilidade de dados apropriados para idiomas não ingleses. O modelo de tradução automática utilizado foi o da Comissão Europeia³, que foi comparado em termos de qualidade de tradução

²<https://opennmt.net/>

³<https://commission.europa.eu/about-european-commission/departments-and-executive>

com o modelo da *Google*. Além disso, os modelos de classificação usados foram o *XLNet*, *RoBERTa* e *RoBERTa*. O estudo visa melhorar a precisão na análise de sentimentos e usa métricas como *Average Recall*, *Average* e *F1-Score* para avaliar o desempenho. Os resultados indicaram que o uso de classificadores monolíngues obtém os melhores resultados na análise de sentimentos para o idioma inglês. Todavia, quando se trata do idioma francês, os classificadores multilíngues demonstraram um bom desempenho, sugerindo que podem ser uma opção viável em idiomas com dados de treinamento limitados.

Com o objetivo de superar as limitações decorrentes da falta de dados em idiomas menos comuns, como turco, espanhol, holandês e russo, no contexto da análise de sentimentos, o estudo realizado por Can, Ezen-Can e Can (2018) apresenta uma abordagem inovadora. A estratégia central envolve a utilização de um classificador de sentimentos previamente treinado em inglês, com o propósito não apenas de contornar os desafios decorrentes da disponibilidade limitada de dados, mas também de ampliar as possibilidades de aplicação desse classificador. Para alcançar esse objetivo, adotou-se uma tática que consiste na tradução dos conjuntos de dados originários dos idiomas mencionados para o inglês, fazendo uso do modelo de tradução automática *SAS Deep Learning Toolkit*. Além disso, o estudo explorou o uso do classificador previamente treinado com conjuntos de dados de avaliações para realizar a classificação dos dados traduzidos. A métrica avaliativa utilizada no experimento foi a acurácia. Os resultados obtidos demonstraram-se relevantes quando comparados à linha de base do modelo RNN, indicando que a tradução automática pode ser um recurso valioso para a coleta de dados em outros idiomas e, assim, melhorar a aplicabilidade de classificadores de sentimentos.

Nos últimos anos, o crescimento das redes sociais desempenhou um papel crucial na área de análise de sentimentos. Essa disciplina, que requer conjuntos de dados substanciais e diversos, tem se beneficiado do acesso facilitado a dados provenientes das redes sociais. Em Balahur e Turchi (2012), foi introduzida a análise de sentimentos em três idiomas distintos: francês, alemão e espanhol. Para isso, foram empregados sistemas de tradução automática, incluindo Bing, Google e Moses. Os resultados obtidos foram considerados satisfatórios e confiáveis baseados na métrica de avaliação *F1 measure*, evidenciando que os sentimentos expressos nos textos se mantêm após a tradução. Esse êxito demonstra a viabilidade de adquirir conjuntos de dados em idiomas além do inglês para treinamento, essa abordagem amplia as perspectivas da análise de sentimentos ao permitir a incorporação de informações de diversas línguas.

Reis et al. (2015) propôs, em seu trabalho, o uso da tradução automática como solução para contornar a falta de recursos na análise de sentimentos em idiomas diferentes do inglês. O processo de tradução automática para o inglês foi realizado com o auxílio

agencies/translation_en

da biblioteca *Python Goslate API*⁴. Foram empregados 13 modelos de análise de sentimentos, que incluíam LIWC, SentiStrength, SentiWordNet, SenticNet, SASA, Happiness Index, PANAS-t, NRC Emotion Lexicon, NRC Hashtag Sentiment Lexicon, Sentiment140 Lexicon e OpinionLexicon, além de Emoticons. Para validar os experimentos, as métricas de desempenho avaliativas utilizadas foram Revocação, Precisão e a *Medida-F1*. Os resultados indicaram que a tradução dos conjuntos de dados não afetou negativamente a acurácia, tornando a tradução viável, desde que seja utilizada uma abordagem de tradução robusta.

A Tabela 2 resume os principais aspectos dos trabalhos relacionados. Observa-se que os conjuntos de dados de entrada são bastante semelhantes, compreendendo principalmente *reviews* de produtos, notícias ou *tweets*. Além disso, as métricas avaliativas empregadas nesses estudos são consistentes, incluindo Precisão, Cobertura e *F1-Score*. Vale ressaltar que grande parte desses trabalhos se concentram na tradução para o inglês, em virtude a gama de técnicas para este idioma.

⁴<http://pythonhosted.org/goslate/>

Trabalho	Modelo de Tradução	Dados	Métricas Avaliativas
“The impact of indirect machine translation on sentiment classification”, (PONCE-LAS et al., 2020)	OpenNMT	<i>Feedback</i>	Acurácia
“The Impact of Machine Translation on Sentiment Analysis”, (SHALUNTS; BACKFRIED; COMMEIGNES, 2016)	SDL Language Weaver	Notícias Gerais	<i>Aggregation, Averaging e Maximization</i>
“Improving sentiment analysis over non-english tweets using multilingual transformers and automatic translation for data-augmentation”, (BARRIERE; BALAHUR, 2020)	Comissão Europeia	<i>Tweets</i>	<i>Average Recall, Average e F1-Score</i>
“Multilingual Sentiment Analysis: An RNN-Based Framework for Limited”, (CAN; EZEN-CAN; CAN, 2018)	SAS Deep Learning Toolkit	<i>Review</i> de Produtos	Acurácia
“Multilingual sentiment analysis using machine translation?”, (BALAHUR; TURCHI, 2012)	Bing, Google e Mo-ses	Notícias Gerais	<i>F1-measure</i>
“Uma Abordagem Multilíngue para Análise de Sentimentos”, (REIS et al., 2015)	Python Goslate API	Tweets	Revocação, Precisão, Medida—F1
Este Trabalho	M2M-100	<i>Tweets, Reviews</i>	Matriz de Confusão, Precisão, Cobertura, <i>F1-Score</i>

Tabela 2: Trabalhos Relacionados.

3 METODOLOGIA

Neste capítulo, descreveremos em detalhes a metodologia utilizada para conduzir o experimento. Os dados de entrada foram obtidos das bases de dados TweetSentBr (BRUM; NUNES, 2017) e B2W-Reviews01 (REAL; OSHIRO; MAFRA, 2019), utilizamos o modelo de tradução neural M2M-100⁵ (FAN et al., 2021) e a classificação foi realizada com o auxílio do classificador Multilingual Twitter-XLM-roBERTa-base⁶ (BARBIERI; ANKE; CAMACHO-COLLADOS, 2021), este experimento pode ser encontrado em um repositório do *GitHub*⁷.

Neste contexto, demonstraremos como as etapas do experimento foram executadas, desde a seleção das bases de dados até a aplicação do modelo de tradução e classificação multilíngue, como é apresentado no fluxograma da Figura 3.

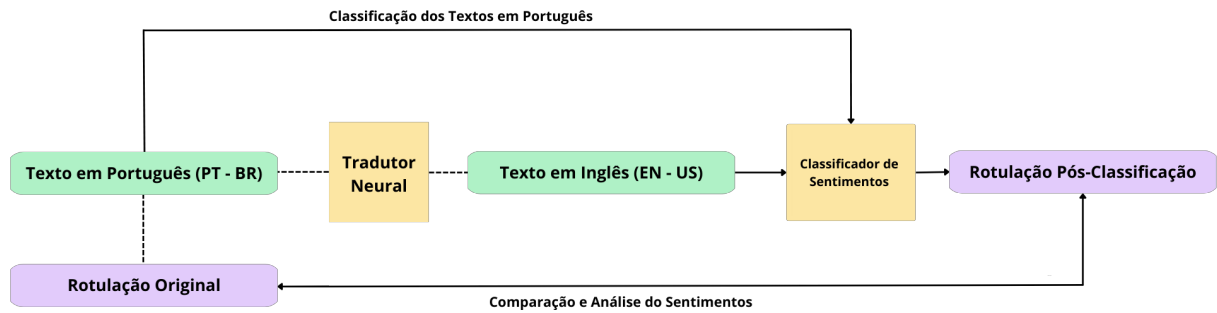


Figura 3: Fluxograma do Experimento

Fonte: Autoria Própria.

3.1 Coleta das Bases de Dados

Os dados utilizados neste estudo foram obtidos por meio de pesquisas no Google Scholar⁸, utilizando termos-chave como “Sentiment Analysis”, “Corpus Brazilian”, e “Sentiment Analysis Corpus Brazilian”. Foram selecionados conjuntos de dados em língua portuguesa que já estavam previamente rotulados. Especificamente, o estudo se concentra em duas dessas bases de dados: TweetSentBr (BRUM; NUNES, 2017) e B2W-Review01 (REAL; OSHIRO; MAFRA, 2019).

3.1.1 TweetSentBr

O conjunto de dados TweetSentBr (BRUM; NUNES, 2017) consiste em 15.000 exemplos de tweets coletados usando a biblioteca Python-Twitter⁹, durante o primeiro

⁵<https://pypi.org/project/dl-translate/>

⁶<https://huggingface.co/cardiffnlp/twitter-xlm-roberta-base-sentiment>

⁷<https://github.com/Lincolnwallace1/Impacto-da-tradu-o-neural-na-classifica-o-de-sentimentos>

⁸<https://scholar.google.com/>

⁹<https://github.com/bear/python-twitter>

semestre de 2017. Essa base de dados foi dividida em conjuntos de treinamento e teste pelo seus autores. O conjunto de treinamento contém 12.999 registros, com rótulos distribuídos da seguinte forma: 44% são positivos, 26% são neutros e 29% são negativos. O conjunto de teste é composto por 2.001 registros, também com uma distribuição semelhante de rótulos: 45% positivos, 25% neutros e 29% negativos. A Tabela 3 mostra exemplos contidos no conjunto de dados.

<i>Tweet</i>	Classe
Que coisa linda! O Programa #encontro estava mostrando uma família que adotou um adolescente de 18 anos. Que amor!!!	positivo
quem viu aquela lutadora modelo barbuda tatuada? #MasterChefBR	neutro
Imagina que insuportável ter de dar de comer pra uma gente que calcula CADA CALORIA que come? Jesus... #MasterChefBR	negativo

Tabela 3: Exemplos de *Tweets* do Conjunto de Dados TweetSentBR.

3.1.2 B2W-Review01

O conjunto de dados B2W-Review01 (REAL; OSHIRO; MAFRA, 2019) originalmente consiste em 132.373 avaliações de produtos coletadas do site Americanas.com. Essas avaliações são classificadas em uma escala de 1 a 5, onde 1 representa uma avaliação ruim e 5 uma avaliação excelente. No entanto, para o propósito deste experimento, foi necessário realizar alguns tratamentos nos dados. As avaliações classificadas com 4 e 5 estrelas foram rotuladas como positivas, enquanto aquelas com 1 e 2 estrelas foram rotuladas como negativas. Vale ressaltar que as avaliações com 3 estrelas não receberam uma classificação específica, sendo assim não foram utilizadas no experimento. Esse pré-processamento dos dados foi realizado especificamente para o estudo de classificadores de avaliações. Esses dados processados foram obtidos por meio do Kaggle¹⁰. Na Tabela 4 são mostrados exemplos de *reviews* contidos no conjunto de dados.

<i>Review</i>	Classe
Bem macio e felpudo...recomendo. Preço imbatível e entrega rápida. Compraria outro quando precisar	positivo
Tive azar, pois o produto não ligou. Tive que entregar para o fornecedor.	negativo

Tabela 4: Exemplos de *Reviews* do Conjunto de Dados B2W.

3.2 Seleção de Ferramenta de Tradução Automática

As traduções dos textos foram realizadas em um ambiente virtual de desenvolvimento, mais precisamente no Colab¹¹, utilizando a linguagem de programação Python¹².

¹⁰<https://www.kaggle.com/datasets/fredericods/ptbr-sentiment-analysis-datasets>

¹¹<https://colab.google/>

¹²<https://www.python.org/>

Para a tradução neural, foi escolhido o modelo M2M-100. Esse modelo foi treinado para lidar com 100 idiomas diferentes, o que permitiu a realização das traduções sem a necessidade de treinar um modelo específico para cada idioma.

Essa escolha do modelo M2M-100 reflete a eficiência da tradução neural, uma vez que ele é capaz de lidar com uma ampla variedade de idiomas sem a necessidade de treinamento adicional. Isso é fundamental para a realização do experimento de análise de sentimentos em diferentes idiomas, pois economiza tempo e recursos.

3.3 Tradução dos Textos

Os textos presentes nos conjuntos de dados não passaram por nenhum tipo de pré-processamento, o que significa que os dados originais podem conter erros ortográficos. Isso ocorre porque esses conjuntos de dados são compostos por tweets e análises de produtos, os quais podem conter linguagem informal e variações ortográficas.

O processo começa com o carregamento dos conjuntos de dados usando a biblioteca *pandas*, que permite a criação de um *DataFrame*. Em seguida, o modelo de tradução “*dlt_translater*” é carregado e configurado para receber textos em português, como entrada, e gerar textos em inglês, como saída. Essa configuração é implementada por meio da função “*translaterText*”, que possui esse propósito.

3.4 Classificações de Sentimentos

Para realizar a classificação dos textos, tanto os originais quanto os traduzidos, foi utilizado o classificador multilíngue *Twitter-XLM-roBERTa-base*. Este classificador foi treinado com um amplo conjunto de dados composto por 198 milhões de *tweets* e foi ajustado para reconhecer 8 idiomas distintos, incluindo o português e o inglês.

É relevante observar que o modelo de classificação, *Twitter-XLM-roBERTa-base*, é derivado da arquitetura *XLM-RoBERTa*, que, por sua vez, é uma otimização do modelo *BERT*. Essa arquitetura, baseada em *transformers*, destaca-se por seu desempenho em tarefas relacionadas ao processamento de linguagem natural.

A escolha desse classificador multilíngue teve como propósito a comparação entre as classificações dos textos antes e depois da tradução. Essa abordagem permitiu avaliar a eficácia das traduções no contexto da análise de sentimentos, ajudando a determinar se a mudança de idioma afetaria ou não a classificação das frases.

3.5 Experimentos

Os experimentos foram divididos em dois estudos distintos, cada um deles utilizando uma base de dados diferente. Essa abordagem visa possibilitar uma análise minu-

cialosa do comportamento das frases em contextos diferentes.

No primeiro experimento, foram utilizados os exemplos de *tweets* do conjunto de dados “TweetSentBr”, conforme apresentado na subseção 3.1.1. No entanto, durante o momento da coleta de dados, foram recuperados apenas 11.524 exemplos dos dados de treinamento, e esses exemplos foram os únicos usados no experimento. Além disso, é importante salientar que esses dados não foram submetidos a nenhum tipo de pré-processamento.

No segundo experimento, foram empregados os exemplos de *reviews* de produtos do conjunto de dados “B2W-Review01”, como detalhado na subseção 3.1.2. Nesse cenário, uma abordagem mais restrita foi adotada, utilizando apenas uma pequena amostra dos dados. Foram selecionados 100 exemplos de cada classe, ou seja, avaliações positivas e negativas. É relevante notar que, devido à origem desses dados ser de outro experimento e terem sido pré-processados por outra pessoa, a classe neutra não foi incluída na seleção. Essa decisão foi tomada em consonância com as escolhas anteriores na utilização desses dados em experimentos prévios. Cabe ressaltar que esses dados foram utilizados sem passar por nenhum pré-processamento adicional antes de serem incorporados aos experimentos.

3.6 Avaliação de desempenho

Para a avaliação comparativa dos resultados, foram utilizadas métricas específicas, o *F1-Score*, a Precisão, Cobertura e a Matriz de Confusão. Todas essas métricas foram calculadas por meio de funções disponíveis na biblioteca Scikit-Learn¹³.

¹³https://scikit-learn.org/stable/modules/model_evaluation.html/

4 APRESENTAÇÃO E ANÁLISE DOS RESULTADOS

No primeiro experimento, como mencionado na Seção 3.5, foi selecionado para o cálculo das métricas de avaliação o parâmetro *average=weighted*, devido ao conjunto de classes desproporcionais. Os resultados deste experimento, considerando os dados no idioma de origem (português), foram os seguintes: *F1-Score* de 59,08%, Precisão de 60,94% e Cobertura de 59,31%, como apresentado na Tabela 5. A escolha do parâmetro *average=weighted* visa obter a pontuação média ponderada. Isso significa que as classes com mais instâncias verdadeiras têm um peso maior na pontuação média ponderada do que as classes com menos instâncias verdadeiras. O resultado é uma métrica de avaliação que leva em consideração o desempenho do modelo em todas as classes, não apenas naquelas com mais instâncias verdadeiras. Os resultados indicam um desempenho moderado do modelo no idioma de origem com base nas métricas de avaliação.

No que diz respeito aos resultados obtidos no idioma traduzido (inglês), utilizando o mesmo parâmetro *average=weighted*, observamos o seguinte desempenho: *F1-Score* de 53,29%, Precisão de 54,73% e Cobertura de 52,62%, como ilustrado na Tabela 5.

Idioma	<i>F1-Score</i>	Precisão	Cobertura
Origem	59,08%	60,94%	59,31%
Traduzido	53,29%	54,73%	52,62%

Tabela 5: Resultados das Métricas de Avaliação do Conjunto de Dados TweetSentBr.

Esses resultados indicam que, após a tradução, houve uma redução no desempenho do modelo em comparação com o idioma original, conforme avaliado pelas métricas.

No segundo experimento, conforme apresentado na Seção 3.5, foi escolhido o parâmetro “*average=micro*” devido à amostra de dados ser um conjunto igualmente balanceado. Os resultados desse experimento no idioma origem (português) foram os seguintes: *F1-Score* de 83%, uma Precisão de 83% e uma Cobertura de 83%, conforme apresentado na Tabela 6.

A escolha do parâmetro “*average=micro*” se deve ao fato de termos uma amostra de dados equilibrada, o que significa que não é necessário atribuir mais ou menos peso a uma classe específica. Essa escolha reflete a ideia de que todas as classes têm igual importância na avaliação do desempenho do modelo.

Os resultados obtidos para essa amostra limitada foram satisfatórios, em parte devido à menor presença de erros ortográficos nesse tipo de texto. Isso torna a análise de sentimentos favorável, uma vez que um classificador depende de um texto de qualidade para funcionar adequadamente.

Em relação aos resultados obtidos no idioma traduzido (inglês), utilizando o mesmo parâmetro “*average=micro*”, observamos o seguinte desempenho: *F1-Score* de 80%, Precisão de 80% e Cobertura de 80%, como apresentado na Tabela 6.

Idioma	<i>F1-Score</i>	Precisão	Cobertura
Origem	83,00%	83,00%	83,00%
Traduzido	80,05%	80,05%	80,05%

Tabela 6: Resultados das Métricas de Avaliação do Conjunto de Dados B2W-Review01.

Os resultados demonstram uma leve perda de desempenho em comparação com o idioma original, conforme avaliado pelas métricas e como aconteceu anteriormente.

De acordo com as matrizes de confusão geradas a partir do conjunto de dados TweetSentBr, conforme apresentado na Figura 4, é possível observar a diagonal principal onde elas indicam a quantidade de classificações corretas, ou seja, os Verdadeiros Positivos. Os outros valores que compõem a matriz de confusão representam as classificações incorretas feitas pelo modelo. Os Falsos Negativos e Falsos Positivos são casos em que o modelo rotulou de forma errada.

Comparando as duas matrizes geradas, podemos destacar algumas diferenças nos Verdadeiros Positivos (casos corretamente classificados) de cada classe. Observamos que a classe de textos negativos teve a maior redução na quantidade de acertos. Na matriz de confusão do idioma origem, essa classe possui 67,37% de acertos, enquanto na matriz do idioma traduzido, esse valor cai para 50%, representando uma diferença de 17,37%. Por outro lado, a classe neutra foi a menos afetada, com uma diminuição relativamente menor na quantidade de acertos. No idioma origem, a classe neutra apresenta 51,44% de acertos, enquanto no idioma traduzido, esse valor é de 48,98%, resultando em uma diferença de 2,48%. Já a classe positiva sofreu uma perda considerável em termos de acertos, com 61,83% no idioma origem e 56,43% no idioma traduzido, uma diferença de 5,4%.

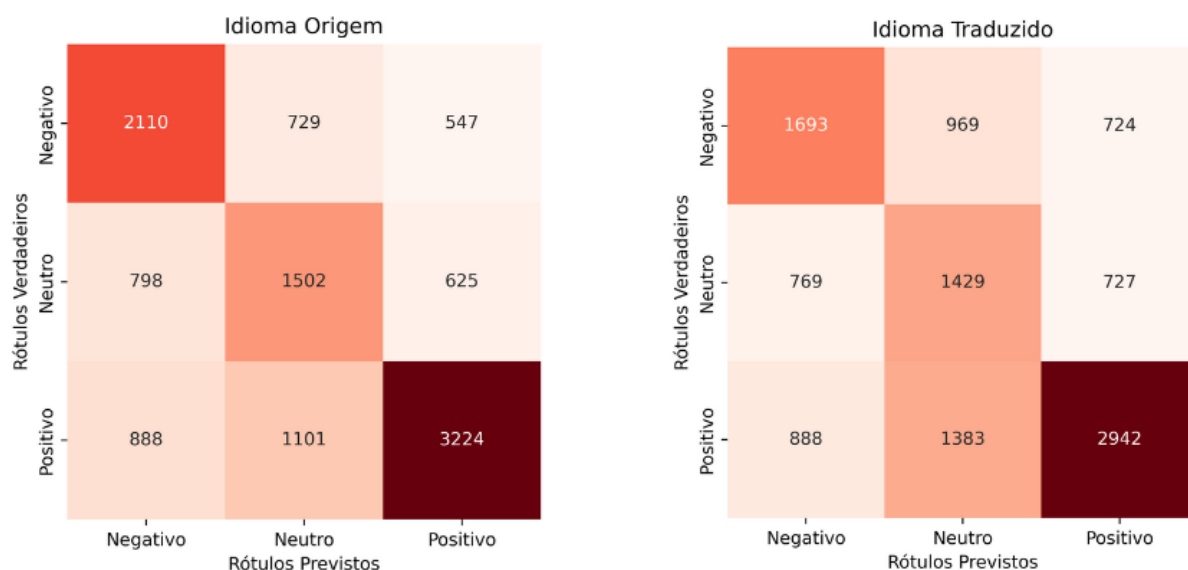


Figura 4: Matrizes de Confusão do Conjunto de Dados TweetSentBr.
Fonte: Autoria Própria

Se observarmos na Tabela 7, os exemplos de *tweets* originalmente rotulados como positivos, entretanto, após a classificação, foram rotuladas como neutras. Esse fenômeno pode ser atribuído ao fato de que essas frases talvez não contenham palavras claramente associadas a sentimentos positivos. Além disso, o contexto da frase como um todo pode não expressar claramente um sentimento positivo, levando, assim, à reclassificação como neutra. Um exemplo disso seria a frase “gosto de sardinha”, na qual a tradução pode alterar o significado original da frase, já que não fica claro se a pessoa gosta de sardinha ou se a comida tem gosto de sardinha e isso sugerir outros sentimentos.

<i>Tweet</i> Original	<i>Tweet</i> Traduzido
Queria um Joaquim Lopes pra mim! #VideoShowAoVivo	I wanted a Joaquim Lopes for me! #VideoShowAoVivo
#MasterChefBR gosto de sardinha	#MasterChefBR likes sardines
”Eu quero esse SAI FOFOCA #VideoShowAovivo”	I want that “SAI FOFOCA” #VideoShowAovivo

Tabela 7: Exemplos de Classificações Neutras Rotuladas Positivas(Falsos Positivos).

Ao analisarmos a Tabela 8, podemos notar alguns exemplos de *tweets* que originalmente foram rotulados como negativos, mas foram classificados como neutros após o processo de análise. Isso pode ser explicado pelo fato de que essas frases não expressam necessariamente sentimentos negativos. Em vez disso, elas podem refletir pensamentos ou gostos pessoais sobre determinado assunto. Dessa forma, essas frases podem não ser verdadeiramente negativas.

Tweet Original	Tweet Traduzido
#Encontro Pensei que estávamos no século XXI Só pensei...	I thought I was in the 21st century and I thought I was...
Só pelo rolê da Mirian hj.. tá demorando! #MasterChefBR	Just for Mirian hj's roll! it's too long! #MasterChefBR
Como giló, mas não como sardinha #MasterChefBr	As a gill, but not as a sardine #MasterChefBr

Tabela 8: Exemplos de Classificações Neutras Rotuladas Negativas(Falsos Positivos).

Nas matrizes de confusão geradas a partir do conjunto de dados B2W-Review01, conforme apresentado na Figura 5, a quantidade de acertos nas classificações dos *reviews* verdadeiros apresenta um aumento quando os textos são traduzidos. No entanto, ocorreu uma pequena perda na quantidade de acertos nas classificações negativas. Além disso, no idioma de origem, os positivos em sua maioria, quando ocorre o erro, é para a classe neutra, diferente do que acontece no idioma traduzido, onde os erros da classe positivo acontecem em sua maioria para a classe negativa. Isso pode ser atribuído ao fato de que alguns dos *reviews* traduzidos não expressaram efetivamente sentimentos negativos. É importante ressaltar que as linhas contendo zeros nas matrizes se devem ao fato de que o conjunto de dados não continha nenhum exemplo de *review* classificado como neutro. No entanto, o classificador pode atribuir uma classificação neutra, dado que o mesmo tem capacidade de retornar três tipos de classes: negativo, positivo e neutro.

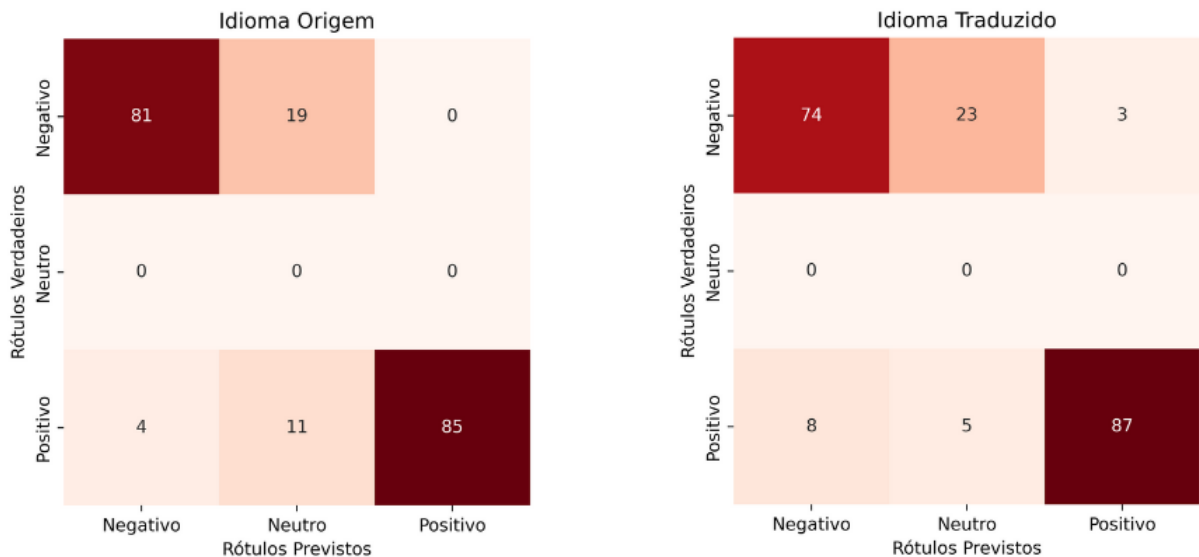


Figura 5: Matrizes de Confusão do Conjunto de Dados B2W-Review01.
Fonte: Autoria Própria

Observando a Tabela 9, com alguns exemplos de *reviews*, fica evidente que o classificador pode estar atribuindo uma classificação neutra devido à falta de expressão de sentimentos positivos. No entanto, quando consideramos o contexto geral do *review*, é

possível perceber que o sentimento positivo realmente está presente.

Rewiew Original	Review Traduzido
já conheço o equipamento, então sei da qualidade do produto	I know the product, so I know the quality of the product.
Indico o aparelho de Ar Condicionado Janela 10.000 Btu's.	I indicate the air conditioner window 10,000 Btu's.
O texto do título fala Gigabit quando o aparelho não é. (10/100)	The title text speaks Gigabit when the device is not. (10/100)

Tabela 9: Exemplos de Classificações Neutras Rotuladas Positivas(Falsos Positivos).

Na Tabela 10, observamos um padrão recorrente nas frases que originalmente foram classificadas como negativas, mas que o classificador rotulou como neutras. Todas essas *reviews* possuem uma palavra em comum: 'não'. O classificador pode estar interpretando a presença dessa palavra como um indicativo de sentimento neutro. Isso pode ocorrer porque o classificador considera o contexto da palavra 'não' como uma negação da frase completa, em vez de avaliar o impacto de uma única palavra.

Rewiew Original	Review Traduzido
Não recebi o produto até agora vou comprar em outro lugar e colocar a americana na justiça já foi pago a primeira parcela	I didn't get the product until now I'll buy elsewhere and put the American in court already paid the first piece
Não era o que eu esperava e aguardo contato para troca.	It wasn't what I expected and I'm waiting for contact for exchange.
Não gruda todos pelos e muito menos qdo vai limpar pra passar de novo, limpa muito poucp	Not grid all the hair and much less qdo will clean to pass again, clean very poucp

Tabela 10: Exemplos de Classificações Neutras Rotuladas Negativas(Falsos Positivos).

É crucial destacar que erros podem ocorrer durante a tradução automática de textos originados no *Twitter*. Esses textos frequentemente contêm erros ortográficos ou sentenças repetidas que podem afetar a qualidade da tradução, como ilustram alguns exemplos de *tweets* na Tabela 11, que foram extraídos do nosso experimento.

Esses problemas de tradução podem levar a uma diminuição no desempenho das classificações. No entanto, é importante salientar que esses problemas podem ser evitados por meio de uma revisão humana mais rigorosa. Muito embora a tradução automática seja uma ferramenta valiosa, ela não é infalível e requer supervisão humana para garantir a qualidade da tradução, especialmente em textos com características específicas, como os mencionados.

Essa ênfase realça a importância de adotar uma abordagem equilibrada entre automação e intervenção humana na análise de sentimentos multilíngue.

5 CONCLUSÕES E TRABALHOS FUTUROS

Com o objetivo de superar a escassez de recursos em um idioma específico, foi proposta a técnica de tradução neural para traduzir textos nesse idioma. Em seguida, aplicamos um classificador de sentimentos multilíngue para avaliar se a tradução neural afeta a análise de sentimentos. Essa abordagem permitiria o uso de tradutores neurais para idiomas com recursos limitados, uma vez que os textos poderiam ser traduzidos para o inglês, ampliando assim as possibilidades de estudo na análise de sentimentos.

Com base nos resultados do primeiro experimento apresentado no Capítulo 4, podemos concluir que a tradução neural se mostra uma opção viável para a coleta de recursos em idiomas específicos. Ela se mostrou satisfatória, uma vez que as métricas avaliativas apresentaram apenas uma perda de eficiência de 6% quando comparamos os experimentos do idioma origem com o idioma traduzido. É importante destacar que, neste problema multi classe, a presença da classe neutra pode influenciar nos resultados, já que as demais classes, quando classificadas incorretamente, tendem a ser rotuladas como neutras.

No segundo experimento podemos, perceber que tivemos apenas uma perda de 3% na eficiência de classificação, isto pode se dar ao fato de ser uma amostra menor e com textos melhores a serem trabalhados, visto que por se tratar de *reviews* esses textos tendem a ter menos erros ortográficos e palavras repetidas.

Reis et al. (2015) apontaram em seus resultados que a tradução neural não prejudica a análise de sentimentos. Portanto, esse achado corrobora com a literatura que afirma que a tradução neural pode ser uma ferramenta viável para a análise de sentimentos, possibilitando a aplicação de modelos de classificação de sentimentos em idiomas com recursos limitados. No entanto, é importante ressaltar que a qualidade da tradução desempenha um papel crucial nesse processo, uma vez que traduções imprecisas podem afetar a precisão da análise de sentimentos. Portanto, ao utilizar a tradução para análise de sentimentos em diferentes idiomas, é essencial garantir que a tradução seja precisa e adequada ao contexto.

Conforme discutido anteriormente, a precisão das traduções é um aspecto crítico para a classificação de sentimentos, sendo essencial para manter a contextualização adequada. Em futuras investigações, sugere-se explorar uma variedade de modelos de tradutores neurais e conjuntos de dados alternativos, considerando a expansão da aplicação da tradução para diferentes gêneros textuais, como notícias gerais, como uma área promissora para pesquisas adicionais.

É relevante observar que ao utilizar um classificador pré-treinado com ampla cobertura de idiomas, sua aplicação em um experimento de duas classes pode apresentar limitações, uma vez que o modelo foi originalmente treinado para reconhecer múltiplas

classes. Como perspectiva futura, o treinamento de um classificador específico para diferentes idiomas e classes pode ser explorado como uma alternativa para melhor se adequar às particularidades de cada língua. Além disso, pode ser considerada a possibilidade de retreinar as últimas camadas do classificador para adaptá-lo a problemas de duas classes.

Essencialmente, a qualidade das traduções desempenha um papel crucial, pois traduções imprecisas podem comprometer negativamente a análise de sentimentos. A escolha criteriosa do modelo de tradução neural e a verificação constante da qualidade das traduções são aspectos críticos. Dessa forma, pesquisas futuras podem se concentrar no desenvolvimento de métodos de avaliação e aprimoramento da tradução para análise de sentimentos em diversos idiomas.

REFERÊNCIAS

- BAHDANAU, D.; CHO, K.; BENGIO, Y. Neural machine translation by jointly learning to align and translate. **arXiv preprint arXiv:1409.0473**, 2014.
- BALAHUR, A.; TURCHI, M. Multilingual sentiment analysis using machine translation? In: **Proceedings of the 3rd workshop in computational approaches to subjectivity and sentiment analysis**. [S.l.: s.n.], 2012. p. 52–60.
- BARBIERI, F.; ANKE, L. E.; CAMACHO-COLLADOS, J. Xlm-t: Multilingual language models in twitter for sentiment analysis and beyond. **arXiv preprint arXiv:2104.12250**, 2021.
- BARRIERE, V.; BALAHUR, A. Improving sentiment analysis over non-english tweets using multilingual transformers and automatic translation for data-augmentation. **arXiv preprint arXiv:2010.03486**, 2020.
- BENEVENUTO, F.; RIBEIRO, F.; ARAÚJO, M. Métodos para análise de sentimentos em mídias sociais. **Sociedade Brasileira de Computação**, 2015.
- BRUM, H. B.; NUNES, M. d. G. V. Building a sentiment corpus of tweets in brazilian portuguese. **arXiv preprint arXiv:1712.08917**, 2017.
- CAN, E. F.; EZEN-CAN, A.; CAN, F. Multilingual sentiment analysis: An rnn-based framework for limited data. **CoRR**, abs/1806.04511, 2018. Disponível em: <<http://arxiv.org/abs/1806.04511>>.
- CHATZIKOUMI, E. How to evaluate machine translation: A review of automated and human metrics. **Natural Language Engineering**, Cambridge University Press, v. 26, n. 2, p. 137–161, 2020.
- DONG, D.; WU, H.; HE, W.; YU, D.; WANG, H. Multi-task learning for multiple language translation. In: **Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**. [S.l.: s.n.], 2015. p. 1723–1732.
- FAN, A.; BHOSALE, S.; SCHWENK, H.; MA, Z.; EL-KISHKY, A.; GOYAL, S.; BAINES, M.; CELEBI, O.; WENZKE, G.; CHAUDHARY, V. et al. Beyond english-centric multilingual machine translation. **The Journal of Machine Learning Research**, JMLRORG, v. 22, n. 1, p. 4839–4886, 2021.
- FRANÇA, T. C. de; OLIVEIRA, J. Análise de sentimento de tweets relacionados aos protestos que ocorreram no brasil entre junho e agosto de 2013. In: SBC. **Anais do III Brazilian Workshop on Social Network Analysis and Mining**. [S.l.], 2014. p. 128–139.
- HIRSCHBERG, J.; MANNING, C. D. Advances in natural language processing. **Science**, American Association for the Advancement of Science, v. 349, n. 6245, p. 261–266, 2015.

- HOSSIN, M.; SULAIMAN, M. N. A review on evaluation metrics for data classification evaluations. **International journal of data mining & knowledge management process**, Academy & Industry Research Collaboration Center (AIRCC), v. 5, n. 2, p. 1, 2015.
- LIDDY, E. D. Natural language processing. 2001.
- LUNARDI, A.; VITERBO, J.; BERNARDINI, F. Análise de sentimentos utilizando técnicas de classificação multiclasse. **Sociedade Brasileira de Computação**, 2016.
- MARTIN, D. J. . J. H. **Speech and Language Processing**. [S.l.]: Computational Linguistics, and Speech Recognition. Speech and Language Processing, 2023.
- PEREIRA, D. A. A survey of sentiment analysis in the portuguese language. **Artificial Intelligence Review**, Springer, v. 54, n. 2, p. 1087–1115, 2021.
- PONCELAS, A.; LOHAR, P.; WAY, A.; HADLEY, J. The impact of indirect machine translation on sentiment classification. **arXiv preprint arXiv:2008.11257**, 2020.
- REAL, L.; OSHIRO, M.; MAFRA, A. B2w-reviews01-an open product reviews corpus. In: **the Proceedings of the XII Symposium in Information and Human Language Technology**. [S.l.: s.n.], 2019. p. 200–208.
- REIS, J.; GONÇALVES, P.; ARAÚJO, M.; PEREIRA, A. C.; BENEVENUTO, F. Uma abordagem multilíngue para análise de sentimentos. In: SBC. **Anais do IV Brazilian Workshop on Social Network Analysis and Mining**. [S.l.], 2015.
- ROCHA, V.; FERNANDES, A.; AGUIAR, S. de. Análise de sentimento sobre filmes no contexto do twitter. In: SBC. **Anais da X Escola Regional de Informática de Mato Grosso**. [S.l.], 2019. p. 115–117.
- SHALUNTS, G.; BACKFRIED, G.; COMMEIGNES, N. The impact of machine translation on sentiment analysis. In: . [s.n.], 2016. Disponível em: <<https://api.semanticscholar.org/CorpusID:36224483>>.
- SUTER, J.; NOGUEIRA, R.; TOZZI, T.; ANDERLE, D.; SPERONI, R. Um data warehouse textual em língua portuguesa: Estudo de caso do sentimento dos usuários do twitter durante a eleição de 2018. In: SBC. **Anais da XV Escola Regional de Banco de Dados**. [S.l.], 2019. p. 41–50.
- TAN, Z.; WANG, S.; YANG, Z.; CHEN, G.; HUANG, X.; SUN, M.; LIU, Y. Neural machine translation: A review of methods, resources, and tools. **AI Open**, Elsevier, v. 1, p. 5–21, 2020.
- VINODHINI, G.; CHANDRASEKARAN, R. Sentiment analysis and opinion mining: a survey. **International Journal**, v. 2, n. 6, p. 282–292, 2012.
- WANG, H.; WU, H.; HE, Z.; HUANG, L.; CHURCH, K. W. Progress in machine translation. **Engineering**, Elsevier, 2021.