Cross-Domain Sentiment Analysis in Portuguese using BERT

Larissa F. S. Britto[1,2], **Luis A. S. Pessoa**[1], **Silvania C. C. Agostinho**[1]

[1]Centro de Pesquisa e Desenvolvimento em Telecomunicações (CPQD)
Campinas – SP – Brazil

[2]Centro de Informática
Universidade Federal de Pernambuco (UFPE) – Recife, PE – Brasil

{lbritto,luisp,scaetano}@cpqd.com.br

Abstract. *Cross-Domain Classification have become a common approach to address the labelled data scarcity problem in Sentiment Analysis (SA). However, the SA domain-dependence and the discrepancy among domains may have a negative impact in classifiers performance. In this work, we evaluated BERT model generalization capability in Cross-Domain Polarity Classification task in Portuguese. For comparison purposes, traditional Machine Learning classifiers and feature extraction approaches are also evaluated. BERT has shown promising results even with the change of domain, achieving 92% of accuracy.*

Resumo. *O Cruzamento de Domínios tem se tornado uma abordagem comum para lidar com a escassez de dados rotulados na Análise de Sentimentos (AS). No entanto, a dependência de domínio da AS e as particularidades associadas a cada domínio podem impactar, negativamente, o desempenho dos modelos de classificação. Neste trabalho, avaliamos a capacidade de generalização do modelo BERT na tarefa de Classificação de Polaridade com Cruzamento de Domínios em Português. Para fins de comparação, classificadores tradicionais de Aprendizagem de Máquina e métodos para extração de características são analisados. O BERT apresentou resultados promissores mesmo com a mudança de domínio, chegando a alcançar 92% de acurácia.*

1. Introdução

Com a popularização da Internet, e-commerces têm se tornado um dos principais meios de compra de produtos. O Brasil é um dos países no qual usuários passam mais tempo na Internet, e também um dos países que mais realiza compras on-line [DataReportal 2022]. Devido às restrições da pandemia de COVID-19, 13 milhões de brasileiros fizeram sua primeira compra pela Internet em 2020 [Neotrust 2022]. Estima-se que em 2021 as vendas on-line no país tenham aumentado 27% em relação ao ano anterior, batendo o recorde anual de faturamento, com um total de mais de R$ 161 bilhões movimentados.

O processo de decisão de compra on-line é repleto de particularidades e envolve diversos fatores, tais como características do produto, avaliação dos usuários, e experiência prévia com a plataforma de vendas. A opinião dos clientes, passada de "boca-a-boca", é um dos principais meios de divulgação de uma empresa, ajudando a criar uma imagem respeitável, e conferindo destaque entre a concorrência [Hossain et al. 2017]. As resenhas (comentários feitos por usuários sobre um determinado produto ou serviço) são

um dos fatores que mais influenciam a decisão de compra na Internet, ultrapassando outros fatores importantes, como recomendações feitas por amigos e experiência prévia do usuário com a plataforma de vendas [DataReportal 2022, Canvas8 and Trustpilot 2020, Insights 2021]. É estimado que 91% dos usuários leiam pelo menos um comentário sobre um determinado produto antes de comprá-lo [Insights 2021]. Por esses motivos, as resenhas de produtos têm se tornado uma forma moderna de boca-a-boca, tendo um papel importante no e-commerce [Chen and Xie 2008].

A disponibilização de dados tão significativos sobre a opinião dos usuários faz da Análise de Sentimentos (AS) uma das subáreas mais populares do Processamento de Linguagem Natural (PLN). A AS tem como objetivo detectar e explorar os sentimentos e/ou opiniões expressos em um documento textual. As empresas utilizam a AS para compreender a opinião de seu público, realizar pesquisas de mercado, avaliar a reputação de marcas e analisar a experiência do cliente, além de obter informações que podem contribuir com a melhoria do produto [Liu et al. 2020].

A principal tarefa da AS é a Classificação de Polaridade, que visa identificar a orientação dos sentimentos ou opiniões em um determinado texto. Essa tarefa consiste da aplicação de técnicas de Aprendizagem de Máquina para classificar documentos textuais, usando características extraídas através de técnicas de PLN. Na Classificação de Polaridade, os documentos são comumente rotulados como *negativos* ou *positivos*, e em alguns casos, a classe *neutra* também é considerada, na qual nenhuma opinião ou sentimentos são expressos no texto [Brum and das Graças Volpe Nunes 2017].

A Classificação de Polaridade possui dependência de uma grande quantidade de dados rotulados para o processo de treinamento dos classificadores. No entanto, em algumas aplicações, a quantidade de dados rotulados pode ser insuficiente, o que impossibilitaria o processo de treinamento e o bom desempenho dos modelos [Abdulraheem et al. 2015]. Para lidar com esse problema, se tornou comum a construção de sistemas em que modelos são treinados com dados de domínios diferentes do domínio de aplicação.

O domínio é um tópico em comum a qual se refere uma coleção de dados, geralmente associados a uma aplicação específica. O domínio reflete em diversas características sobre os dados. Na AS, o estilo de escrita pode variar de um domínio para outro, e palavras importantes em um domínio podem não existir em outro, como pode ser observado no exemplo da Figura 1. Na Classificação de Polaridade com Cruzamento de Domínios, os modelos são treinados com exemplos de um problema rico em dados (*domínio de origem*) e testados com documentos de outro problema, no qual há escassez de dados (*domínio de destino*) [Peng et al. 2018], portanto, exige um maior nível de generalização por parte dos classificadores.

Neste trabalho, analisamos a performance do modelo BERT para o cruzamento de domínios na Classificação de Polaridade de sentimentos em bases de dados em Português. O modelo é comparado com classificadores tradicionais de Aprendizagem de Máquina: Regressão Logística, Naive Bayes, Floresta Aleatória e Máquinas de Vetores de Suporte. Além disso, duas abordagens populares para extração de características também são comparadas: GloVe [Pennington et al. 2014] e TF-IDF [Salton and McGill 1983]. Essa comparação tem como objetivo analisar a capacidade de generalização de cada um

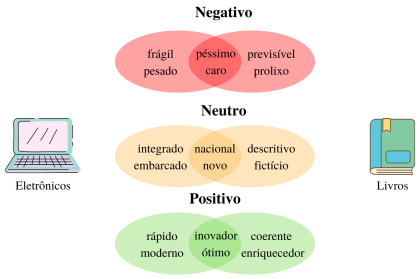


Figura 1. Exemplo de como termos importantes de cada polaridade podem variar de domínio a domínio. É possível observar que, apesar da ocorrência de termos comuns, existem termos exclusivos de cada domínio.

dos modelos, a fim de compreender qual seria mais adequado para compor um sistema real com aplicação em domínios diversos.

O trabalho está dividido como segue. Na próxima seção (Seção 2), alguns trabalhos relacionados serão brevemente discutidos. Em seguida, é apresentada uma breve descrição das bases de dados utilizadas (Seção 3), dos métodos para extração de características (Seção 4), e dos classificadores (Seção 5) selecionados para a avaliação experimental. Na Seção 6, os resultados experimentais serão discutidos. Por fim, na Seção 7, as conclusões do trabalho são apresentadas.

2. Trabalhos Relacionados

Nos últimos anos, esforços têm sido investidos em diversas tarefas da AS no Português, com o intuito de fornecer recursos para o desenvolvimento de pesquisas e aplicações nesta área. Entre as tarefas tratadas, podemos citar a detecção de discurso de ódio [Soto et al. 2019, O. Plath et al. 2022], detecção de ironia e sarcasmo [Schubert and de Freitas 2020, Gonçalves et al. 2015] e a principal delas: classificação de polaridade.

Em [Britto and Pacífico 2019], a classificação de polaridade é executada numa base de dados composta por resenhas de aplicativos para smartphones. Redes Neurais Recorrentes (modelo tradicional e a variação *Long Short-Term Memory*) são comparadas com diversos classificadores tradicionais de Aprendizagem de Máquina, como Naive Bayes, Árvore de Decisão, Floresta Aleatória, Máquinas de Vetores de Suporte, e Regressão Logística. Os modelos neurais e o classificador Regressão Logística foram responsáveis pelos melhores resultados. O processo de desenvolvimento da base de dados é descrito, permitindo que outros pesquisadores consigam replicá-lo para a construção de suas próprias bases de dados. A base de dados proposta, no entanto, contempla apenas documentos das classes positivas e negativas, não abordando as resenhas neutras.

As resenhas neutras possuem grande importância na AS, tendo em vista que sistemas de aplicação real de AS constantemente recebem documentos neutros como entrada, e esses documentos apresentam diferenças significativas em relação às demais classes (positiva e negativa) [Brum and das Graças Volpe Nunes 2017]. A detecção dos documentos neutros pode, inclusive, melhorar o desempenho dos classificadores na classe positiva e negativa [Koppel and Schler 2006]. Adotando o uso dos documentos neutros na classificação de polaridade, [Brum and das Graças Volpe Nunes 2017] propôs uma base

de dados de *tweets* em Português sobre programas de TV, com cerca de 15 mil sentenças anotadas manualmente. A classificação de polaridade é executada usando os classificadores Naive Bayes e Máquinas de Vetores de Suporte, além de um modelo híbrido, usando uma abordagem baseada em léxicos.

Apesar das abordagens citadas anteriormente apresentarem boa performance, elas dependem da disponibilidade de uma grande quantidade de dados rotulados, e a anotação muitas vezes é um processo custoso e demorado [Júnior et al. 2017]. Alguns trabalhos têm sido desenvolvidos no instuito de lidar com esse problema através do uso de abordagens de cruzamento de domínios. Em [Júnior et al. 2017], uma abordagem de Supervisão de Distância foi utilizada para criar uma grande base de dados de *tweets* rotulados. Essa base de dados é usada no treinamento de modelos baseados em Regressão Logística e Redes Neurais Convolucionais, além de modelos híbridos baseados em léxicos. Os modelos são testados em documentos nos domínios de política e resenhas de produtos do Buscapé e do Mercado Livre.

Em [Britto et al. 2019], a Classificação de Polaridade com Cruzamento de Domínios é tratada com modelos baseados em um método popular para adaptação de domínios: o *Structural Correspondence Learning* [Blitzer et al. 2006]. Para avaliar o modelo, bases de dados de resenhas de produtos de diferentes domínios foram propostas. A avaliação mostrou que o método de adaptação de domínio foi capaz de melhorar a performance do classificador Regressão Logística.

Os trabalhos discutidos que fazem uso da técnica de cruzamento de domínios não contemplam bases de dados com documentos neutros, que, como já citado, possuem um papel fundamental em sistemas de aplicação real. Até onde avaliamos, não há nenhum trabalho que aborde o cruzamento de domínios para AS que contemplem bases de dados em Português que incluam documentos neutros. Visando lidar com essa limitação, este trabalho apresenta as seguintes contribuições:

- Executar a Classificação de Polaridade com Cruzamento de Domínios com **múltiplas classes** (negativo, neutro e positivo) no idioma **Português**.
- Avaliar o desempenho de **diferentes classificadores** da literatura de AS, na tarefa de Classificação de Polaridade com Cruzamento de Domínios.
- Analisar o impacto que **diferentes métodos para extração de características** podem ter nos modelos, na tarefa de Classificação de Polaridade com Cruzamento de Domínios.

3. Base de Dados

Para a avaliação dos modelos adotados neste trabalho, foram selecionadas bases de dados para Análise de Sentimentos em Português, compostas por comentários de usuários/consumidores sobre determinados produtos/serviços, anotadas entre três polaridades: negativo, neutro ou positivo. Foram priorizadas bases de dados anotadas manualmente, no intuito de garantir a qualidade da anotação. Bases de dados dos seguintes domínios foram adotadas:

- **Livros** [Freitas et al. 2014] - Este *corpus* é composto por cerca de 12 mil frases anotadas manualmente, extraídas de 1600 resenhas, de 13 livros, de 7 autores diferentes, com uma média de cerca de 200 resenhas por autor, contemplando diferentes gêneros,

Tabela 1. Estatísticas das bases de dados adotadas.

Base de Dados	Classe	N. de Documentos	Vocabulário	Tamanho Médio (Palavras)	Tamanho Mín - Máx (Palavras)
Livros (origem)	Neg	557	2039	12.79	1 - 70
	Neu	8946	14678	14.12	1 - 178
	Pos	2735	5182	11.96	1 - 90
	Total	12238	16393	13.58	1 - 178
Buscapé (destino)	Neg	81	1177	31.89	1 - 138
	Neu	20	138	8.9	1 - 16
	Pos	815	2059	10.89	1 - 98
	Total	916	2725	12.7	1 - 138
Eletrônicos (destino)	Neg	59	853	27.85	2 - 101
	Neu	43	466	17.35	2 - 96
	Pos	131	1059	19.31	1 - 139
	Total	233	1779	21.11	1 - 139

alcançando assim uma grande variedade de estilos: de escrita muito informal, com uso de gírias, abreviaturas, neologismos e *emoticons*, para resenhas mais formais, com um vocabulário mais refinado. Por possuir um tamanho significativamente maior que as demais bases analisadas, esta base de dados foi selecionada para fazer parte do treinamento dos classificadores (domínio de origem).

- **Produtos Variados** (**Buscapé**) [Hartmann et al. 2014] - Grande conjunto de comentários sobre produtos da plataforma de vendas Buscapé. Como a anotação desta base de dados foi feita de forma automática, contemplando apenas as polaridades positiva e negativa, neste trabalho, a anotação foi revisada de forma manual, visando aumentar a qualidade da base de dados, e detectar os comentários neutros incorretamente anotados em outras classes. Foram selecionados aleatoriamente um conjunto de cerca de mil comentários. Entre os produtos mais frequentes nessas resenhas estão perfumes, brinquedos, jogos, eletrodomésticos e eletrônicos. Por possuir um tamanho relativamente pequeno, o que poderia ser insuficiente para o processo de aprendizagem dos modelos de classificação, esta base de dados foi utilizada apenas para o teste dos algoritmos (domínio de destino).
- Produtos **Eletrônicos** [Belisário et al. 2020] - *Corpus* contendo cerca de 230 comentários sobre produtos eletrônicos do site Buscapé (sem sobreposição com a base de dados de produtos variados, citada anteriormente). O *corpus*, anotado manualmente, foi originalmente utilizado para classificação de subjetividade em Português. Assim como a base de dados ***Buscapé***, esta base de dados também foi adotada para teste dos classificadores, devido à pouca quantidade de dados.

Na abordagem adotada neste trabalho, a base de dados de origem foi dividida em treino e teste, numa proporção 90%-10%, respectivamente. Para evitar que o desbalanceamento das classes, presente em todas as bases de dados, prejudicasse o processo de aprendizagem, criando um viés nos classificadores, foi empregada uma etapa de balanceamento nos dados de treino, na qual o número de documentos das classes majoritárias foi reduzido para o número de documentos da classe minoritária, igualando assim a quantidade de dados de cada classe. Em todas as bases de dados foram ainda removidos os comentários repetidos, ou que não continham textos (que continham apenas números, pontuações ou símbolos). Algumas informações sobre as bases de dados adotadas podem ser vistas na Tabela 1. Além disso, os termos mais frequentes de cada uma das bases de dados podem ser vistos na Figura 2.

Apesar dos domínios apresentarem muitos termos em comum, como “ótimo”, “bom”, “péssimo”e “interessante”, é possível observar como cada domínio possui suas

(a) Livros (b) Buscapé (c) Eletrônicos

Figura 2. Termos mais frequentes de cada domínio. Para facilitar a visualização, foi feita a remoção de stopwords.

particulares através de termos bem específicos, como “prende”, “trama”e “narrativa”(no domínio de ***Livros***), “perfurme”, “fragância”, “fixação”, “durabilidade”, “consumo”(no domínio de produtos do ***Buscapé***) e “assistência”, “desempenho”e “configuração”(no domínio de ***Eletrônicos***).

4. Extração de Características

Uma das etapas fundamentais da classificação de texto é a extração de características, através da qual é possível transformar dados textuais em matrizes numéricas suportadas pelos modelos de classificação. Os seguintes métodos de extração de características foram adotados:

- ***Global Vectors for Word Representation*** (**GloVe**) [Pennington et al. 2014] é um dos algoritmos mais populares de aprendizado não supervisionado para a representação de documentos textuais através de *word embeddings* [1]. O GloVe combina as vantagens de dois tipos de técnicas frequentemente adotadas na literatura: fatoração de matriz global e janelas de contexto local [Pennington et al. 2014]. Uma das vantagens do modelo é ser treinado apenas nos elementos diferentes de zero na matriz de coocorrência de palavras, em vez de em toda a matriz esparsa.
- ***Term Frequency-Inverse Document Frequency*** (**TF-IDF**) [Salton and McGill 1983] é um método estatístico comumente utilizado para extração de características em textos. O TF-IDF (Equação 3) combina o TF (Equação 1), a proporção do número de vezes que a palavra aparece em um documento, com o IDF (Equação 2), que tem como objetivo mensurar o quão importante um termo é, no intuito de diminuir a influência de termos que ocorrem com uma grande frequência, mas que possuem pouca relevância.

$$TF_{t,d}=\frac{f_{t,d}}{\sum_{t_n \in d} f_{t_n,d}} \quad (1) \qquad IDF_{t,d}=\log \frac{N}{df_t} \quad (2) \qquad TF{-}IDF_{t,d}=TF_{t,d}\times IDF_{t,d} \quad (3)$$

na qual a função f retorna a frequência do termo t no documento d, N é o número total de documentos, df é o número de documentos em que t ocorre.

5. Classificadores

Nesta seção descrevemos brevemente os classificadores usados neste trabalho.

- **Regressão Logística (*Logistic Regression* - LR)** - é um modelo estatístico que visa prever a probabilidade dos possíveis resultados de uma variável dependente dado um conjunto de variáveis independentes, assumindo que a variável dependente pode ser prevista pela combinação linear de características do problema e parâmetros do modelo.

[1] https://paperswithcode.com/method/glove

- **Naive Bayes (NB)** é dos modelos mais populares e simples para classificação de texto. O modelo probabilístico é baseado no teorema de Bayes, e usa a probabilidade de ocorrência de cada evento para prever a classe do documento. O modelo assume que todas as características são independentes umas das outras dada a classe, ignorando qualquer relação entre elas.
- **Floresta Aleatória (*Random Forest* - RF)** é um método que combina um conjunto de Árvores de Decisão (estrutura similar a fluxogramas, que realiza predições baseado em regras de decisão simples). O uso de um conjunto de Árvores de Decisão visa evitar os efeitos que ruídos e *outliers* podem ter na saída de uma única árvore, resultando em um classificador muito mais robusto. O classificador seleciona a classe com mais votos sobre todas as árvores da floresta [Criminisi et al. 2011].
- **Máquinas de Vetores de Suporte (*Support Vector Machines* - SVM)** são algoritmos de aprendizado supervisionado, utilizados para mapear o espaço de características de entrada em um novo espaço, no qual as classes são linearmente separáveis [Bergsma et al. 2005]. Para fazer isso, o SVM tem como objetivo encontrar um hiperplano que pode separar melhor as instâncias de diferentes classes.
- **Representações de Codificador Bidirecional de Transformadores (*Bidirectional Encoder Representations from Transformers* - BERT)** [Devlin et al. 2018] é um método neural simples e poderoso, que apesar de recente, já é um dos modelos mais populares para várias tarefas de PLN. O BERT foi projetado para pré-treinar representações bidirecionais profundas de textos não rotulados. Enquanto os modelos anteriores com o mesmo propósito treinam o modelo na ordem da sequência das palavras, o BERT usa duas direções, analisando tanto o contexto à esquerda quanto à direita da palavra.

6. Avaliação Experimental e Resultados

Nesta seção, os resultados experimentais serão apresentados e discutidos. Os experimentos visam analisar a performance de técnicas de Aprendizagem de Máquina para a classificação de polaridade de sentimentos com cruzamento de domínios. Todos os modelos foram treinados na base de dados no domínio de ***Livros***, e testados tanto nesse domínio, quanto nos demais domínios selecionados (***Buscapé*** e ***Eletrônicos***). Diferentes métodos para extração de característica (GloVe e TF-IDF) e classificação (Regressão Logística, Naive Bayes, Floresta Aleatória, Máquinas de Vetores de Suporte e BERT) serão comparados.

O *framework HuggingFace (Transformers)*[2] foi utilizado na implementação do classificador BERT, adotando o modelo BERTimbau [Souza et al. 2020] (BERT pré-treinado em grande *corpus* em Português). Uma análise empírica nos valores de hiperparâmetros sugeridos em [Devlin et al. 2018] foi realizada, através da qual não foi possível observar grande impacto no desempenho do classificador com a variação dos hiperparâmetros testados. Os seguintes valores foram adotados: o tamanho do *batch* (*batch_size*) foi definido como 32 e a taxa de aprendizagem (*learning_rate*) como 2e-5. O modelo foi ajustado (*fine-tuned*) por 4 épocas, e a cada época o modelo gerado foi armazenado, sendo selecionado o modelo para o qual a menor taxa de erro no conjunto de validação foi obtida. Os demais classificadores foram implementa-

[2] `https://huggingface.co/`

dos através da biblioteca para Aprendizagem de Máquina *Scikit-Learn* [3]. Os hiperparâmetros desses classificadores foram otimizados através do *framework Optuna* [4], tomando como espaço de busca valores comumente utilizados na literatura de AS [Shekhar et al. 2022, Yang and Shami 2020]. Um processo de Validação Cruzada com 5-*folds* é adotada no conjunto de treino, visando evitar resultados obtidos por sorte. Os valores de hiperparâmetros dos classificadores com maior Acurácia (Equação 5) média são selecionados.

Para a avaliação dos experimentos, algumas das métricas mais utilizadas para classificação de texto foram adotadas: Acurácia, Revocação (*Recall*), Precisão e *F-Measure* (F1), as fórmulas dessas métricas podem ser vistas abaixo (Equações 5 - 7)

$$Recall = \frac{TP}{TP+FN} \quad (4) \qquad Acurácia = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

$$Precisão = \frac{TP}{TP+FP} \quad (6) \qquad F1 = \frac{2 \times Precisão \times Recall}{Precisão + Recall} \quad (7)$$

onde, TP e TN equivalem a qantidade de documentos positivos e negativos classificados corretamente, e FN e FP a quantidade de documentos positivos e negativos, respectivamente, categorizados de forma errônea.

Os resultados obtidos podem ser visualizados na Tabela 2. Ao avaliar a influência das características na performance dos classificadores, foi possível observar que, em determinadas situações, as características GloVe foram responsáveis por resultados superiores, como ao serem testadas na base de dados de ***Livros***, com os classificadores Regressão Logística e Naive Bayes. No entanto, o fenômeno oposto também foi observado em outros cenários, nos quais o TF-IDF apresentou uma melhoria de até de 15% na acurácia, ao ser testado na base de dados ***Buscapé*** com o classificador Floresta Aleatória. Na maioria dos cenários testados, o TF-IDF representou uma melhoria no desempenho dos classificadores, no entanto, os resultados obtidos demonstram a necessidade de uma análise aprofundada (com um maior volume de dados e variedade de domínios) para definir quais características, de fato, seriam mais adequadas para os classificadores em um sistema real de AS com cruzamento de domínios.

Através da Tabela 2, podemos observar ainda que, assim como esperado, a mudança de domínio prejudicou o desempenho dos modelos, em particular, dos classificadores de Aprendizagem Rasa, nos quais no melhor modelo (TF-IDF + SVM) ocorreu uma perda de até 20% de acurácia do domínio de origem para o domínio de produtos ***Eletrônicos***. No melhor caso (TF-IDF + NB), os classificadores rasos alcançaram 55.4% de acurácia no domínio de ***Eletrônicos***, um valor baixo se comparado com o melhor desempenho no domínio original (TF-IDF + LR com 66.5% acurácia). Ao contrário dos modelos tradicionais de Aprendizagem de Máquina, o BERT foi capaz de superar o problema de mudança de domínio, mantendo resultados significativos em todos os testes, alcançando 78% de acurácia na base original, e 92% e 74% de acurácia na base ***Buscapé*** e ***Eletrônicos***, respectivamente.

Apesar da alta taxa de acerto, ao analisar as demais métricas adotadas, é possível observar valores relativamente baixos, como no domínio de ***Livros***, no qual a precisão no modelo com melhor desempenho, o BERT, não ultrapassou 30%. Esse fenômeno já era

[3] https://scikit-learn.org/

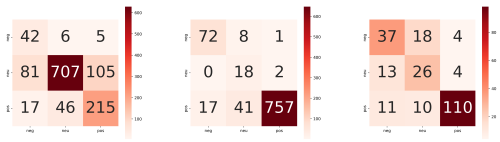
[4] https://optuna.org/

Tabela 2. Resultados experimentais por classe.

Modelos		LR		NB		RF		SVM		BERT
Características		GloVe	TF-IDF	GloVe	TF-IDF	GloVe	TF-IDF	GloVe	TF-IDF	
Livros → Livros										
Acurácia	Neg	0.8342	0.8382	0.8105	0.7525	0.7802	0.8178	0.8350	0.8374	**0.9109**
	Neu	0.7165	0.6928	0.6356	0.6168	0.6577	0.6471	0.7067	0.7092	**0.8056**
	Pos	0.7794	0.7598	0.7467	0.7386	0.7582	0.7361	0.7721	0.7786	**0.8587**
	-	0.6650	0.6454	0.5964	0.5539	0.5980	0.6005	0.6569	0.6626	**0.7876**
Precisão	Neg	0.1560	0.1596	0.0987	0.1166	0.0970	0.1336	0.1502	0.1557	**0.3000**
	Neu	0.8679	0.8752	0.8301	0.9061	0.8475	0.8700	0.8648	0.8755	**0.9315**
	Pos	0.5152	0.4752	0.4506	0.4441	0.4672	0.4390	0.4982	0.5118	**0.6615**
	Macro	0.5130	0.5033	0.4598	0.4889	0.4706	0.4809	0.5044	0.5143	**0.631**
Recall	Neg	0.6415	0.6415	0.4151	0.7170	0.4906	0.5849	0.6038	0.6226	**0.7925**
	Neu	0.7212	0.6753	0.6293	0.5297	0.6473	0.6069	0.7088	0.7010	**0.7917**
	Pos	0.4892	0.5504	0.5252	0.6007	0.4604	0.5827	0.5000	0.5468	**0.7734**
	Macro	0.6173	0.6224	0.5232	0.6158	0.5328	0.5915	0.6042	0.6235	**0.7858**
F1	Neg	0.2509	0.2556	0.1594	0.2005	0.1620	0.2175	0.2406	0.2491	**0.4352**
	Neu	0.7878	0.7623	0.7159	0.6686	0.7340	0.7150	0.7791	0.7786	**0.8559**
	Pos	0.5018	0.5100	0.4850	0.5107	0.4638	0.5008	0.4991	0.5287	**0.7131**
	Macro	0.5135	0.5093	0.4535	0.4599	0.4532	0.4778	0.5063	0.5188	**0.6681**
Tempo Treino		0.3365	0.4354	0.2272	**0.0389**	2.9143	0.3326	0.6242	0.5146	94.2313
Tempo de Teste		0.0384	0.0063	0.0353	**0.0066**	0.0516	0.0304	0.1154	0.0602	3.3835
Livros → Buscapé										
Acurácia	Neg	0.8155	0.8897	0.8406	0.8581	0.6823	0.8275	0.7380	0.8734	**0.9716**
	Neu	0.8068	0.7707	0.8493	0.9127	0.8111	0.8275	0.7937	0.7369	**0.9443**
	Pos	0.7096	0.7107	0.7751	0.8013	0.5677	0.7074	0.6124	0.6692	**0.9334**
	-	0.6659	0.6856	0.7325	0.7860	0.5306	0.6812	0.5721	0.6397	**0.9247**
Precisão	Neg	0.2442	0.4180	0.2593	0.3356	0.1635	0.2825	0.1858	0.3723	**0.809**
	Neu	0.0514	0.0519	0.0597	0.0714	0.0629	0.0400	0.0576	0.0526	**0.2687**
	Pos	0.9824	0.9725	0.9706	0.9541	0.9883	0.9643	0.9873	0.9812	**0.9961**
	Macro	0.4260	0.4808	0.4299	0.4537	0.4049	0.4289	0.4102	0.4687	**0.6912**
Recall	Neg	0.5185	0.6296	0.4321	0.6173	0.6296	0.6173	0.5802	0.6296	**0.8889**
	Neu	0.4500	0.5500	0.4000	0.2500	0.5500	0.3000	0.5500	0.6500	**0.9000**
	Pos	0.6859	0.6945	0.7706	0.8160	0.5202	0.6969	0.5718	0.6405	**0.9288**
	Macro	0.5515	0.6247	0.5342	0.5611	0.5666	0.5381	0.5673	0.6400	**0.9059**
F1	Neg	0.3320	0.5025	0.3241	0.4348	0.2595	0.3876	0.2814	0.4679	**0.8471**
	Neu	0.0923	0.0948	0.1039	0.1111	0.1128	0.0706	0.1043	0.0974	**0.4138**
	Pos	0.8078	0.8103	0.8591	0.8796	0.6817	0.8091	0.7242	0.7751	**0.9613**
	Macro	0.4107	0.4692	0.4290	0.4752	0.3513	0.4224	0.3700	0.4468	**0.7407**
Tempo de Teste		0.0848	**0.0155**	0.0865	0.0171	0.1187	0.0559	0.4197	0.1959	7.0007
Livros → Eletrônicos										
Acurácia	Neg	0.6652	0.6781	0.6781	0.6524	0.5708	0.6524	0.6223	0.6695	**0.8026**
	Neu	0.6266	0.6738	0.6996	0.7682	0.6438	0.6781	0.6695	0.6352	**0.8069**
	Pos	0.6352	0.7039	0.6953	0.6867	0.5665	0.6567	0.6009	0.6738	**0.8755**
	-	0.4635	0.5279	0.5365	0.5536	0.3906	0.4936	0.4464	0.4893	**0.7425**
Precisão	Neg	0.3538	0.3621	0.3519	0.3553	0.2796	0.3553	0.3253	0.3200	**0.6066**
	Neu	0.2800	0.3187	0.3247	0.3721	0.2727	0.2895	0.3191	0.2941	**0.4815**
	Pos	0.8382	0.8690	0.7941	0.7544	0.7885	0.8148	0.8393	0.8395	**0.9322**
	Macro	0.4907	0.5166	0.4902	0.4939	0.4469	0.4865	0.4946	0.4845	**0.6734**
Recall	Neg	0.3898	0.3559	0.3220	0.4576	0.4407	0.4576	0.4576	0.2712	**0.6271**
	Neu	0.6512	0.6744	0.5814	0.3721	0.5581	0.5116	0.6977	0.6977	**0.6047**
	Pos	0.4351	0.5573	0.6183	0.6565	0.3130	0.5038	0.3588	0.5191	**0.8397**
	Macro	0.4920	0.5292	0.5072	0.4954	0.4373	0.4910	0.5047	0.4960	**0.6905**
F1	Neg	0.3710	0.3590	0.3363	0.4000	0.3421	0.4000	0.3803	0.2936	**0.6167**
	Neu	0.3916	0.4328	0.4167	0.3721	0.3664	0.3697	0.4380	0.4138	**0.5361**
	Pos	0.5729	0.6791	0.6953	0.7020	0.4481	0.6226	0.5027	0.6415	**0.8835**
	Macro	0.4451	0.4903	0.4827	0.4914	0.3855	0.4641	0.4403	0.4496	**0.6788**
Tempo de Teste		0.0384	**0.0063**	0.0353	0.0066	0.0516	0.0304	0.1154	0.0602	03.3835

esperado, e não significa má performance dos modelos avaliados, e sim uma consequência do desbalanceamento da base de dados de teste, tendo em vista que esses valores baixos só são observados nas classes minoritárias (positivo e negativo no dominio de ***Livros***, neutro no domínio ***Buscapé*** e negativo e neutro no domínio de ***Eletrônicos***). O número de falsos positivos da classe majoritária, apesar de baixo, tem um grande peso na precisão das classes minoritárias (como pode ser visto na Equação 6). Esse fenômeno pode ser comprovado através da análise das matrizes de confusão (Figura 3).

O BERT apresentou um tempo de treino de 94 segundos, alto quando comparado com os demais modelos, que levaram no máximo 4 segundos no treinamento. No entanto, o tempo de teste (fator determinante para que um modelo possa compor um sistema real) alcançou no pior dos casos 7 segundos, o que pode ser considerado um tempo aceitável, quando comparado ao tempo e performance dos demais modelos.



(a) Livros (b) Buscapé (c) Eletrônicos

Figura 3. Matriz de Confusão do modelo BERT para cada domínio. Através do mapa de calor é possível observar o alto peso que as classes majoritárias apresentam.

7. Conclusões

Neste trabalho, diversos modelos de classificação foram analisados, com o objetivo de avaliar a capacidade generalização desses modelos na Classificação de Polaridade de Sentimentos com Cruzamento de Domínios. Esta comparação possibilitaria a escolha do modelo mais adequado para compor um sistema de aplicação real em domínios diversos.

Apesar de ser um modelo simples e antigo para extração de características, o TF-IDF foi capaz de superar o GloVe, um modelo recente e popular, nos cenários testados. Os classificadores tradicionais de Aprendizagem de Máquina Rasa foram altamente impactados pela mudança de domínio, ao contrário do BERT, que foi capaz de obter os melhores resultados, chegando a alcançar uma diferença de até 20% de acurácia em comparação com os demais classificadores, sendo capaz também de manter bons resultados, mesmo com a mudança de domínio, alcançando 92% de acurácia quando temos ***Livros*** como domínio de origem, e ***Buscapé*** como domínio alvo.

Como trabalhos futuros, pretendemos tratar, através do desenvolvimento de novas bases de dados e revisão de bases já existentes, uma das principais limitações do nosso trabalho, o pequeno número de documentos, especificamente na classe neutra, uma vez que a maior parte das bases de dados da literatura não contempla esta classe. Também pretendemos analisar como etapas de pré-processamento de texto podem impactar o resultado dos classificadores, assim como testar outros modelos populares na literatura da Análise de Sentimentos.

Referências

Abdulraheem, A., Abdullah Arshah, R., and Qin, H. (2015). Evaluating the Effect of Dataset Size on Predictive Model Using Supervised Learning Technique. *International Journal of Software Engineering Computer Sciences (IJSECS)*, 1:75–84.

Belisário, L. B., Ferreira, L. G., and Pardo, T. A. S. (2020). Evaluating Methods of Different Paradigms for Subjectivity Classification in Portuguese. In Quaresma, P., Vieira, R., Aluísio, S., Moniz, H., Batista, F., and Gonçalves, T., editors, *Computational Processing of the Portuguese Language*, pages 261–269, Cham. Springer International Publishing.

Bergsma, S., Jung, D., Lau, R., Wang, Y., and Wang, S. (2005). Machine learning approaches to sentiment classification cmput 551 : Course project winter , 2005.

Blitzer, J., McDonald, R., and Pereira, F. (2006). Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 Conference on Empirical Methods*

in Natural Language Processing, EMNLP '06, page 120–128, USA. Association for Computational Linguistics.

Britto, L. and Pacífico, L. (2019). Análise de sentimentos para revisões de aplicativos mobile em português brasileiro. In *Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional*, pages 1080–1090, Porto Alegre, RS, Brasil. SBC.

Britto, L. F., Lima, R., and Pacífico, L. D. S. (2019). Structural correspondence learning for cross-domain sentiment analysis in brazilian portuguese. In *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 812–817.

Brum, H. B. and das Graças Volpe Nunes, M. (2017). Building a sentiment corpus of tweets in brazilian portuguese. *CoRR*, abs/1712.08917.

Canvas8 and Trustpilot (2020). The critical role of reviews in internet trust. `https://business.trustpilot.com/guides-reports/build-trusted-brand/the-critical-role-of-reviews-in-internet-trust`. (Accessed on 07/09/2022).

Chen, Y. and Xie, J. (2008). Online consumer review: Word-of-mouth as a new element of marketing communication mix. *Management Science*, 54(3):477–491.

Criminisi, A., Konukoglu, E., and Shotton, J. (2011). Decision forests for classification, regression, density estimation, manifold learning and semi-supervised learning.

DataReportal (2022). Digital 2022 global digital overview. `https://datareportal.com/reports/digital-2022-global-overview-report`. (Accessed on 07/09/2022).

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Freitas, C., Motta, E., Milidiú, R., and César, J. (2014). *Sparkling Vampire... lol! Annotating Opinions in a Book Review Corpus*, pages 128–146.

Gonçalves, P., Dalip, D., Reis, J., Messias, J., Ribeiro, F., Melo, P., Araújo, L., Gonçalves, M., and Benevenuto, F. (2015). Bazinga! caracterizando e detectando sarcasmo e ironia no twitter. In *Anais do IV Brazilian Workshop on Social Network Analysis and Mining*, page , Porto Alegre, RS, Brasil. SBC.

Hartmann, N., Avanço, L., Balage, P., Duran, M., das Graças Volpe Nunes, M., Pardo, T., and Aluísio, S. (2014). A large corpus of product reviews in Portuguese: Tackling out-of-vocabulary words. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3865–3871, Reykjavik, Iceland. European Language Resources Association (ELRA).

Hossain, M. M., Kabir, S., and Rezvi, R. (2017). Influence of Word of Mouth on Consumer Buying Decision: Evidence from Bangladesh Market. *International Journal of Business and Management*, 9:38–45.

Insights, B. (2021). Customer reviews' impact on purchase decisions — bizrate insights. `https://bizrateinsights.com/resources/shopper-survey-report-the-impact-reviews-have-on-consumers-purchase-decisions/`. (Accessed on 07/09/2022).

Júnior, E. A. C., Marinho, V. Q., dos Santos, L. B., Bertaglia, T. F. C., Treviso, M. V., and Brum, H. B. (2017). Pelesent: Cross-domain polarity classification using distant supervision. *CoRR*, abs/1707.02657.

Koppel, M. and Schler, J. (2006). The Importance of Neutral Examples for Learning Sentiment. *Computational Intelligence*, 22(2):100–109.

Liu, H., Chatterjee, I., Zhou, M., Lu, X. S., and Abusorrah, A. (2020). Aspect-based sentiment analysis: A survey of deep learning methods. *IEEE Transactions on Computational Social Systems*, 7(6):1358–1375.

Neotrust (2022). Com pandemia, vendas pela internet crescem 27% e atingem r$ 161 bi em 2021. `https://neotrust.com.br/2022/04/08/e-commerce-brasileiro-tem-melhor-faturamento-dos-ultimos-anos-em-janeiro/`. (Accessed on 07/09/2022).

O. Plath, H., O. Paiva, M. E., L. Pinto, D., and D. P. Costa, P. (2022). Detecção de Discurso de Ódio Contra Mulheres em Textos em Português Brasileiro: Construção da Base MINA-BR e Modelo de Classificação. *Revista Eletrônica de Iniciação Científica em Computação*, 20(3 SE - Edição Especial: CTIC/CSBC).

Peng, M., Jiang, Y.-g., and Huang, X. (2018). Cross-Domain Sentiment Classification with Target Domain Specific Information. pages 2505–2513.

Pennington, J., Socher, R., and Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Salton, G. and McGill, M. (1983). *Introduction to Modern Information Retrieval*. International student edition. McGraw-Hill.

Schubert, G. and de Freitas, L. (2020). A construção de um corpus para detecção de ironia e sarcasmo em português. In *Anais do XVII Encontro Nacional de Inteligência Artificial e Computacional*, pages 709–717, Porto Alegre, RS, Brasil. SBC.

Shekhar, S., Bansode, A., and Salim, A. (2022). A comparative study of hyper-parameter optimization tools. *CoRR*, abs/2201.06433.

Soto, C., Nunes, G., and Gomes, J. (2019). Avaliação de técnicas de word embedding na tarefa de detecção de discurso de ódio. In *Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional*, pages 1020–1031, Porto Alegre, RS, Brasil. SBC.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: Pretrained bert models for brazilian portuguese. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I*, page 403–417, Berlin, Heidelberg. Springer-Verlag.

Yang, L. and Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *CoRR*, abs/2007.15745.