Exploring model transfer strategies for sentiment analysis in Twitter

Eliseu Guimarães[1,2], Jonnathan Carvalho[3], Aline Paes[1], Alexandre Plastino[1]

[1]Instituto de Computação – Universidade Federal Fluminense – Brazil

[2]Marinha do Brasil – Brazil

[3]Instituto Federal Fluminense – Brazil

eliseuguimaraes@id.uff.br, joncarv@iff.edu.br

{alinepaes,plastino}@ic.uff.br

Abstract. *Social media have become trendy environments for communication. Because of that, analyze the sentiment that the user expresses in their social media posts is an important research field. However, detecting polarity in such contents is a challenge, partially because the amount of labeled data to train classifiers is scarce in many situations. This paper explores strategies for reusing a model learned from a source dataset to classify instances in a target dataset. The experiments are conducted with 22 tweets sentiment analysis datasets and approaches based on similarity metrics. The results point out that the size of the source training set plays an essential role in the classifiers' performance when they were applied to the target data.*

Resumo. *As mídias sociais se tornaram um ambiente popular para comunicação. Por isso, analisar o sentimento que o usuário expressa em suas postagens nas redes sociais é um importante campo de pesquisa. No entanto, detectar a polaridade em tais conteúdos é um desafio, em parte porque a quantidade de dados rotulados para treinar classificadores é escassa em muitas situações. Este artigo explora estratégias para reusar um modelo aprendido a partir de conjunto de dados fonte para classificar instâncias em um conjunto de dados de destino. Os experimentos são conduzidos com 22 conjuntos de dados de análise de sentimento em tweets e abordagens baseadas em métricas de similaridade. Os resultados apontam que o tamanho do conjunto de treinamento fonte desempenha um papel essencial no desempenho dos classificadores quando usados para inferir a classe das instâncias alvo.*

1. Introdução

A análise de sentimentos consiste no estudo computacional de identificar opiniões, sentimentos, emoções, humores e atitudes das pessoas [Liu 2020]. Com o surgimento e popularização das redes sociais, como o Twitter[1], qualquer pessoa pode expressar livremente suas opiniões e sentimentos a respeito de assuntos variados através de textos curtos, os tweets. Tweets são considerados um desafio para a análise de sentimentos devido às

[1]http://www.twitter.com

suas características, como uso incorreto da gramática, presença frequente de erros ortográficos, falta de contexto devido à limitação a apenas 280 caracteres, bem como a presença de sarcasmo, ironia subjetividade, entre outros [Martínez-Cámara et al. 2014].

A detecção de polaridade em tweets – objetivo deste estudo – consiste na classificação das opiniões expressas em tweets quanto às suas polaridades, aqui tratadas como positivas ou negativas. Abordagens comumente utilizadas para tratar este problema se baseiam em técnicas de aprendizado de máquina, que visam extrair características de tweets previamente rotulados em um dado domínio para treinar classificadores capazes de determinar a polaridade de novos tweets naquele mesmo domínio [Barbosa and Feng 2010, Dong et al. 2014]. Contudo, nem sempre é possível obter dados rotulados em quantidade suficiente para o treinamento de classificadores que alcancem um bom desempenho preditivo. Isso pode ocorrer tanto pela escassez de dados do domínio de interesse, quanto pelo esforço necessário para rotular manualmente uma grande quantidade de dados, muitas vezes proibitivo.

Uma possível solução para o problema da escassez de dados rotulados em um domínio de interesse é aproveitar um classificador aprendido anteriormente para a mesma tarefa, e *adaptá-lo* ou *reusá-lo* no domínio pretendido. Essas soluções são investigadas na área de transferência de aprendizado [Pan and Yang 2010], uma vez que as instâncias do domínio de interesse alvo, em geral, são amostradas a partir de uma distribuição distinta da que originou o domínio fonte de treinamento. Entretanto, mesmo oriundos de distribuições distintas, podem existir conjuntos de dados que são mais promissores para a transferência do que outros. No entanto, a seleção adequada de tais conjuntos de dados é um problema desafiador, para o qual diversos estudos têm sido conduzidos na literatura [Guimarães et al. 2020, Guo et al. 2018, Li et al. 2017, Ruder and Plank 2017, Santos et al. 2019]. Esses estudos incluem a avaliação do uso de métricas de similaridade para selecionar uma base de dados apropriada do domínio-fonte (base-fonte) ou selecionar um subconjunto de instâncias do domínio-fonte para treinar um classificador para a base no domínio-alvo (base-alvo), entre outras investigações.

O trabalho apresentado em [Guimarães et al. 2020] visa selecionar a base-fonte mais apropriada para treinar um classificador para uma determinada base-alvo, no contexto da análise de sentimentos em tweets. Nesse caso, a base-fonte é selecionada por meio da análise de similaridade entre a base-alvo e cada base-fonte candidata avaliada, a partir de um conjunto pré-definido de 21 bases-fonte candidatas rotuladas. Dessa forma, a base-fonte candidata mais similar à base-alvo é selecionada para treinar um classificador, que em seguida é avaliado de acordo com suas habilidades preditivas na base-alvo. Em [Guimarães et al. 2020], são avaliadas quatro métricas de similaridade, mas, no geral, os classificadores com melhor poder preditivo foram aqueles treinados com bases-fonte selecionadas por meio da similaridade de cosseno.

Apesar dos resultados promissores obtidos em [Guimarães et al. 2020], não foi investigada a hipótese da utilização de todos os dados disponíveis, por meio da união de todas as bases-fonte candidatas, por exemplo, para treinar o classificador. Além disso, também não foram exploradas estratégias para selecionar um subconjunto de instâncias das bases-fonte, utilizando critérios de similaridade e dissimilaridade entre as instâncias.

Neste contexto, este artigo apresenta um conjunto de experimentos computaci-

onais com o objetivo de responder às questões de pesquisa definidas a seguir: **Q1** – *Dado um conjunto de bases-fonte rotuladas, vale a pena selecionar uma delas para treinar um classificador de polaridade para uma base-alvo não-rotulada, como feito em [Guimarães et al. 2020], ou um melhor desempenho preditivo poderia ser obtido se o classificador de polaridade para a base-alvo fosse treinado a partir da união de todas as bases-fonte disponíveis?* e **Q2** – *Considerando a união de todas as bases-fonte disponíveis, vale a pena selecionar um subconjunto de suas instâncias com base na similaridade em relação às instâncias da base-alvo?* Os resultados obtidos indicam que unir diferentes bases-fonte candidatas para compor um conjunto de treinamento mais amplo gera classificadores com maior poder preditivo do que os treinados a partir da seleção de uma única base-fonte.

O restante deste artigo está organizado da seguinte forma. A Seção 2 apresenta os trabalhos relacionados. Na Seção 3, são descritos os experimentos conduzidos neste estudo para responder às questões de pesquisa Q1 e Q2. A Seção 4 apresenta os resultados e os discute. Por fim, na Seção 5, são apresentadas as conclusões e trabalhos futuros.

2. Trabalhos Relacionados

Além do trabalho apresentado em [Guimarães et al. 2020], diversos outros têm se dedicado ao estudo de técnicas para composição de um conjunto de treinamento a partir de bases-fonte, com o intuito de aprender um classificador para ser aplicado a uma base-alvo [Guo et al. 2018, Li et al. 2017, Ruder and Plank 2017, Santos et al. 2019].

Em [Guo et al. 2018], é proposta uma abordagem do tipo *mixture-of-experts*, na qual se considera que diferentes bases-fonte estão alinhadas a regiões distintas da base-alvo. Uma métrica que relaciona as instâncias da base-alvo com as bases-fonte é aprendida, para ponderar os resultados de classificadores treinados utilizando as bases-fonte. Os resultados do estudo mostram que as acurácias preditivas obtidas usando esse método foram superiores às obtidas utilizando como conjunto de treinamento apenas uma base-fonte ou a união de todas as bases-fonte.

Por sua vez, [Li et al. 2017] consideram que as instâncias da base-alvo que estejam mais próximas da base-fonte têm maior probabilidade de serem corretamente classificadas. Desta forma, são atribuídos pesos maiores a essas instâncias e é utilizada uma regularização para assegurar uma propagação suave dos rótulos na base-alvo. Esta abordagem foi aplicada em um conjunto de 12 pares de bases e os resultados de acurácia obtidos foram comparados com nove abordagens estabelecidas em outros estudos, mostrando que essa técnica obteve a melhor posição no ranqueamento.

A abordagem proposta em [Ruder and Plank 2017] utiliza otimização Bayesiana para aprender uma métrica de similaridade de bases, que é definida como uma combinação linear de um conjunto de atributos. Foram utilizadas seis métricas de similaridade entre bases como atributos para esse aprendizado, calculadas considerando três tipos de representações dos dados, além de seis métricas de diversidade aplicadas ao conjunto de treinamento. Os resultados apresentados mostram que as acurácias utilizando métricas de similaridade combinadas com métricas de diversidade apresentam melhor desempenho do que utilizar apenas similaridade ou apenas diversidade, além de superarem os resultados de compor o conjunto de treinamento com seleção aleatória ou utilizando uma métrica específica de reconhecido desempenho.

Em [Santos et al. 2019], são utilizadas métricas de similaridade para a seleção de bases-fonte em português, com o intuito de treinar classificadores para uma base-alvo de tweets no contexto das eleições presidenciais no Brasil em 2018. Nesse trabalho, são utilizadas abordagens de composição do conjunto de treinamento que incluem mesclar as bases mais semelhantes à base-alvo e mesclar as bases menos semelhantes à base-alvo. Os resultados indicam que utilizar bases-fonte mais semelhantes é vantajoso e, ao mesmo tempo, que a inclusão de bases-fonte menos semelhantes ao conjunto de treinamento deteriora o desempenho do classificador.

Diferentemente do proposto por trabalhos anteriores, este estudo combina: (i) a utilização de uma métrica única para a seleção de dados, o que evita o aprendizado e treinamento de uma métrica; (ii) a seleção de instâncias isoladas, o que permite que a quantidade de instâncias da base-fonte dissimilares à base-alvo seja limitada; (iii) a consideração do uso da dissimilaridade como parte do método de seleção, o que pode ajudar a reduzir o overfitting, trazendo diversidade ao conjunto de treinamento; e (iv) o uso de um amplo conjunto de bases-fonte, o que torna os resultados robustos ao utilizar bases com grande variedade de características.

3. Metodologia Experimental

Nesta seção, a metodologia adotada nos experimentos computacionais reportados neste estudo é detalhada. Na Seção 3.1, é apresentada a configuração dos experimentos computacionais e, na Seção 3.2, são detalhados os experimentos conduzidos para responder às questões de pesquisa Q1 (Seção 3.2.1) e Q2 (Seção 3.2.2).

3.1. Configuração dos Experimentos Computacionais

Nos experimentos computacionais conduzidos neste estudo, são utilizadas 22 bases de dados de tweets em inglês[2] [Carvalho and Plastino 2021]. As bases são compostas de tweets que expressam opiniões sobre diversos assuntos e rotuladas quanto às suas polaridades, ou seja, se são opiniões positivas ou negativas. Quanto ao conteúdo, enquanto algumas bases contêm tweets sobre um tema específico, como *movie* e *hobbit* (filmes), *archeage* (jogos) e *OMD* (política), outras são compostas de tweets com conteúdo mais geral, como *Narr*, *SemEval18* e *Vader*, por exemplo. Além disso, as bases variam em tamanho (quantidade de tweets) e distribuição de classes, como pode ser observado na Tabela 1.

Tabela 1. Quantidade de tweets positivos (#pos) e negativos (#neg), percentual de tweets positivos (%pos) e total de tweets das bases de dados.

Bases de dados	#pos	#neg	%pos	Total	Bases de dados	#pos	#neg	%pos	Total
irony	22	43	34%	65	sarcasm	33	38	46%	71
aisopos	159	119	57%	278	SemEval15-Fig	47	274	15%	321
sentiment140	182	177	51%	359	person	312	127	71%	439
hobbit	354	168	68%	522	iphone	371	161	70%	532
movie	460	101	82%	561	sanders	570	654	47%	1224
Narr	739	488	60%	1227	archeage	724	994	42%	1718
SemEval18	865	994	47%	1859	OMD	710	1196	37%	1906
HCR	539	1369	28%	1908	STS-gold	632	1402	31%	2034
SentiStrength	1340	949	59%	2289	Target-dependent	1734	1733	50%	3467
Vader	2897	1299	69%	4196	SemEval13	3183	1195	73%	4378
SemEval17	2375	3972	37%	6347	SemEval16	8893	3323	73%	12216

[2] https://github.com/joncarv/air-datasets

Como pré-processamento dos tweets, todas as menções a usuários foram substituídas pelo token único *@user* e URLs foram substituídas pelo token *http://www.url.com*. Todos os tweets foram colocados em letras minúsculas e, então, tokenizados. Como atributos para o treinamento dos classificadores é utilizada a abordagem de *word embeddings*. A geração dos *embeddings* dos tweets foi feita calculando, para cada um, a média dos *embeddings* de seus tokens. Para isso, foi adotado o modelo estático pré-treinado apresentado em [Bravo-Marquez et al. 2016]. Este modelo foi treinado em um conjunto de 10 milhões de tweets com o método Skip-gram [Mikolov et al. 2013] e possui 400 dimensões.

O algoritmo de classificação adotado nos experimentos é o de regressão logística, que obteve bom desempenho preditivo no contexto da análise de sentimentos em tweets em [Carvalho and Plastino 2021]. Nesse caso, foi utilizada a implementação da biblioteca scikit-learn[3], com o valor máximo de iterações igual a 10.000, de modo a evitar falhas na convergência do algoritmo. Para a avaliação dos classificadores foram adotadas as medidas acurácia e F_1-*measure* (ponderada).

3.2. Descrição dos Experimentos Computacionais

Esta seção descreve os experimentos conduzidos neste estudo para responder às questões de pesquisa Q1 (Seção 3.2.1) e Q2 (Seção 3.2.2), introduzidas na Seção 1.

3.2.1. Questão de Pesquisa Q1

O experimento descrito nesta seção visa responder à questão de pesquisa, Q1 – *Dado um conjunto de bases-fonte rotuladas, vale a pena selecionar uma delas para treinar um classificador de polaridade para uma base-alvo não-rotulada, como feito em [Guimarães et al. 2020], ou um melhor desempenho preditivo poderia ser obtido se o classificador de polaridade para a base-alvo fosse treinado a partir da união de todas as bases-fonte disponíveis?*

Este experimento investiga a hipótese de utilização de todos os dados disponíveis em bases-fonte candidatas de diversos domínios para treinar um classificador para uma base-alvo, por meio da união das instâncias dessas bases-fonte. Assim, considerando as 22 bases de dados apresentadas na Seção 3.1, cada uma é tratada uma vez como base-alvo e a união das 21 bases restantes é a base-fonte. Dessa forma, um classificador é treinado usando como conjunto de treinamento todos os tweets da base-fonte resultante dessa união. Essa estratégia é denominada **Estratégia 21D**.

Para cada base-alvo avaliada, o classificador treinado com a união das 21 bases restantes é aplicado à base-alvo e o desempenho preditivo (acurácia e F_1) é comparado ao obtido quando o classificador é treinado com a base-fonte candidata mais similar à base-alvo, obtida por meio da similaridade de cosseno, como reportado em [Guimarães et al. 2020] (**Estratégia SC**).

Além disso, o desempenho preditivo também é comparado com aquele obtido pelo classificador treinado com a própria base-alvo, após a execução de uma validação cruzada com 10 partições (**Estratégia Alvo**). No entanto, cabe ressaltar que, para a análise

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

experimental que está sendo conduzida neste estudo, considera-se que a base-alvo é não-rotulada. Esse fato impediria o treinamento de um classificador com a base-alvo. De todo modo, essa situação é considerada como um *baseline* para fins de comparação com o desempenho da estratégia 21D descrita nesta seção.

3.2.2. Questão de Pesquisa Q2

Os experimentos descritos nesta seção visam responder à questão de pesquisa, Q2 – *Considerando a união de todas as bases-fonte disponíveis, vale a pena selecionar um subconjunto de suas instâncias com base na similaridade em relação às instâncias da base-alvo?*

Para estes experimentos, o objetivo é treinar um classificador para a base-alvo utilizando um subconjunto das instâncias do conjunto união das bases-fonte. Nesse caso, para cada base-alvo, o conjunto de bases-fonte é formado pela união das 21 bases-fonte restantes disponíveis. Chamaremos esse conjunto de C_{all} daqui em diante.

Para selecionar as instâncias de C_{all} que comporão o conjunto de treinamento $C_{train} \subset C_{all}$, duas estratégias de seleção de instâncias são investigadas neste estudo, descritas a seguir.

Estratégia S1: Nesta estratégia, C_{train} é composto por um percentual p de instâncias oriundas de C_{all}. São explorados diferentes valores de p, onde $0 < p \leq 100, p \in \mathbb{N}$. Duas formas de seleção são analisadas: (**sim**) *seleção de instâncias similares* – seleção das instâncias mais similares à base-alvo, e (**dis**) *seleção de instâncias similares e dissimilares* – seleção de instâncias mais similares e de instâncias menos similares à base-alvo, em uma razão de 4:1. Por exemplo, com $p = 5$, são selecionadas 4% das instâncias de C_{all} que sejam mais similares à base-alvo e 1% das instâncias menos similares. O cálculo da similaridade é realizado por meio da similaridade de cosseno, devido ao bom desempenho reportado em [Guimarães et al. 2020].

Para computar a similaridade, a base-alvo será representada por uma única representação vetorial de 400 dimensões. Para isso, duas estratégias de representação são analisadas: (**mt**) *média de tokens* – a representação como sendo a média dos *embeddings* de todos os *tokens* que compõem a base, e (**mi**) *média de instâncias* – a representação como sendo a média dos *embeddings* das instâncias pertencentes à base.

Com relação à distribuição de classes dos conjuntos de treinamento C_{train}, duas situações são analisadas: (i) **sem balanceamento** – seleção de instâncias mantendo a distribuição original de classes de C_{all}, e (ii) **com balanceamento** – seleção de instâncias com uma distribuição balanceada. Nesse caso, para a estratégia de seleção *sim*, as instâncias da classe *majoritária menos similares* à base-alvo não são selecionadas. Por outro lado, para a estratégia de seleção *dis*, como são selecionadas as mais similares e as menos similares em uma razão de 4:1, as instâncias intermediárias da classe majoritária não são selecionadas. Por exemplo, se 700 instâncias pertencem à classe majoritária e 500 à classe minoritária, quando $p = 100$, são selecionadas, da classe majoritária, as 400 instâncias mais similares à base-alvo e as 100 menos similares ($400 + 100 = 500$, que é a quantidade de instâncias da classe minoritária).

Estratégia S2: Nesta estratégia, para cada base-alvo avaliada, o conjunto de treinamento C_{train}, é formado selecionando-se, para cada instância da base-alvo, as k instâncias de C_{all}

mais similares a ela. A seleção é feita de maneira iterativa, como descrito a seguir. Na primeira iteração, a instância em C_{all} mais similar a cada instância da base-alvo é selecionada. Na próxima iteração, a segunda instância mais similar a cada instância da base-alvo é selecionada e, assim, sucessivamente até a k-ésima iteração. Este procedimento assegura que, para dois valores i e j, tais que $k_i < k_j$, cada conjunto de treinamento C_{train} gerado obedece à relação $C_{train_i} \subset C_{train_j}$.

Para avaliar as estratégias S1 e S2, os classificadores obtidos a partir dos conjuntos C_{train} são aplicados à base-alvo e o desempenho preditivo (acurácia e F_1) é comparado com aquele obtido pelo classificador treinado com as instâncias da própria base-alvo, após a execução de uma validação cruzada com 10 partições e com os resultados obtidos pela Estratégia 21D.

4. Resultados Experimentais

Nesta seção, são reportados os resultados dos experimentos para responder às questões de pesquisa Q1 (Seção 4.1 e Q2 (Seção 4.2).

4.1. Respondendo à Questão de Pesquisa Q1

A Tabela 2 apresenta os resultados obtidos para responder à questão de pesquisa Q1. A segunda, terceira e quarta colunas apresentam as acurácias preditivas dos classificadores treinados com a própria base-alvo (estratégia Alvo), com a base-fonte selecionada pela similaridade de cosseno (estratégia SC), reportados em [Guimarães et al. 2020], e com a base formada pela união das 21 bases-fonte (estratégia 21D). De forma semelhante, a sétima, oitava e nona colunas apresentam os valores de F_1. Os melhores resultados encontram-se sublinhados. Observando os valores registrados na tabela, a estratégia *21D* apresenta melhores resultados que a estratégia *Alvo* em 16 das 22 bases-alvo em termos de acurácia, e em 17 das 22 bases-alvo em termos de F_1.

Além dos desempenhos preditivos das estratégias avaliadas, também são analisados os ganhos obtidos ao treinar um classificador com determinada estratégia (SC ou 21D), em relação a utilizar a própria base-alvo como conjunto de treinamento. Nesse caso, valores de ganho maiores que 1 significam que treinar um classificador com a estratégia avaliada produz um desempenho melhor do que utilizar a base-alvo.

Na Tabela 2, a quinta coluna apresenta, em termos de acurácia, os valores dos ganhos obtidos comparando o desempenho da estratégia SC [Guimarães et al. 2020] com o desempenho da estratégia Alvo (coluna *SC x Alvo*). A sexta coluna indica os ganhos de acurácia obtidos comparando o desempenho da estrategia 21D com o desempenho da estratégia Alvo (coluna *21D x Alvo*). De forma semelhante, na décima e na décima-primeira colunas, os ganhos apresentados são referentes aos resultados obtidos em termos de F_1. Por fim, em negrito encontram-se assinalados os maiores valores de ganho.

Analisando os ganhos obtidos pelas estratégias avaliadas, é possível observar que treinar um classificador usando a união de todas as bases-fonte candidatas disponíveis (colunas *21D x Alvo*) gera um resultado melhor do que treinar um classificador com uma única base-fonte selecionada pela similaridade de cosseno (colunas *SC x Alvo*), para 20 das 22 bases-alvo, tanto para acurácia quanto para F_1. Além disso, considerando os valores médios dos ganhos obtidos, apresentados na última linha (*Ganho médio*), é possível notar um aumento considerável tanto para acurácia (de 0,92 para 1,03) quanto para F_1

Tabela 2. Análise de desempenho dos classificadores treinados com a união das bases-fonte disponíveis.

Bases de dados	Acurácia					F_1-*measure*				
	Alvo	SC*	21D	Ganho		Alvo	SC*	21D	Ganho	
				SC x Alvo	21D x Alvo				SC x Alvo	21D x Alvo
irony	0,66	0,68	0,77	1,02	**1,16**	0,53	0,68	0,76	1,30	**1,45**
sarcasm	0,56	0,58	0,76	1,02	**1,35**	0,43	0,53	0,76	1,24	**1,78**
aisopos	0,81	0,86	0,88	1,06	**1,08**	0,80	0,86	0,88	1,07	**1,10**
SemEval15-Fig	0,85	0,70	0,59	**0,82**	0,69	0,79	0,74	0,65	**0,94**	0,83
sentiment140	0,81	0,69	0,86	0,85	**1,07**	0,80	0,67	0,86	0,83	**1,07**
person	0,71	0,73	0,81	1,03	**1,13**	0,59	0,71	0,80	1,19	**1,35**
hobbit	0,68	0,69	0,75	1,02	**1,11**	0,55	0,64	0,74	1,16	**1,35**
iphone	0,70	0,71	0,74	1,02	**1,06**	0,57	0,72	0,75	1,26	**1,31**
movie	0,82	0,81	0,83	0,99	**1,01**	0,74	0,78	0,83	1,06	**1,12**
sanders	0,76	0,61	0,77	0,80	**1,01**	0,76	0,57	0,77	0,76	**1,02**
Narr	0,83	0,66	0,89	0,79	**1,07**	0,83	0,64	0,89	0,78	**1,08**
archeage	0,82	0,57	0,77	0,70	**0,93**	0,82	0,54	0,77	0,66	**0,94**
SemEval18	0,77	0,63	0,81	0,82	**1,05**	0,77	0,60	0,81	0,78	**1,05**
OMD	0,76	0,65	0,71	0,85	**0,94**	0,73	0,65	0,69	0,90	**0,95**
HCR	0,72	0,73	0,74	1,01	**1,03**	0,60	0,62	0,67	1,04	**1,11**
STS-gold	0,78	0,80	0,71	**1,03**	0,91	0,75	0,80	0,72	**1,07**	0,97
SentiStrength	0,75	0,71	0,79	0,94	**1,05**	0,74	0,67	0,79	0,90	**1,06**
Target-dependent	0,80	0,66	0,77	0,83	**0,97**	0,80	0,65	0,77	0,82	**0,97**
Vader	0,83	0,81	0,84	0,98	**1,01**	0,81	0,81	0,84	1,00	**1,04**
SemEval13	0,77	0,77	0,83	1,01	**1,08**	0,71	0,73	0,81	1,03	**1,15**
SemEval17	0,85	0,62	0,85	0,72	**0,99**	0,85	0,60	0,85	0,71	**1,00**
SemEval16	0,82	0,80	0,84	0,97	**1,02**	0,81	0,77	0,84	0,96	**1,04**
	Ganho médio:			0,92	**1,03**	Ganho médio:			0,97	**1,13**

*Resultados reportados em [Guimarães et al. 2020]

(de 0,97 para 1,13) ao utilizar a união de todas as bases-fonte candidatas, em detrimento a selecionar uma base-fonte específica.

4.2. Respondendo à Questão de Pesquisa Q2

Esta seção apresenta as avaliações das estratégias de seleção de instâncias, S1 e S2, descritas na Seção 3.2.2, para responder à questão de pesquisa Q2.

4.2.1. Estratégia de Seleção S1

A estratégia S1 consiste na seleção de um percentual p das instâncias da base-fonte. Na Tabela 3, os resultados reportados correspondem aos valores de ganho médio obtidos ao avaliar as possíveis combinações da representação da base-alvo com a forma de seleção de instâncias. Mais especificamente, as formas de representação da base-alvo, *mt* (utilizando a média dos *embeddings* dos *tokens*) e *mi* (utilizando a média dos *embeddings* das instâncias), são combinadas com as formas de seleção de instâncias, *sim* (selecionando apenas as instâncias mais similares à base-alvo) e *dis* (selecionando as mais similares e as mais dissimilares), a saber: ***mt+sim***, ***mt+dis***, ***mi+sim*** e ***mi+dis***. O ganho médio consiste na média dos ganhos para as 22 bases de dados avaliadas. Devido ao espaço limitado, para cada combinação avaliada, são reportados apenas os resultados dos cinco subconjuntos C_{train} que obtiveram os melhores ganhos médios.

A parte esquerda da Tabela 3 apresenta a avaliação do cenário em que o percentual selecionado segue a distribuição original de classes da base-fonte (*sem balanceamento*) e na parte direita estão os resultados obtidos quando são selecionadas quantidades iguais de instâncias positivas e negativas (*com balanceamento*). Os resultados em negrito indicam os casos em que os ganhos médios são superiores aos obtidos ao utilizar toda a base-fonte, como reportado na Tabela 2 (1,03 e 1,13, em termos de acurácia e F_1, respectivamente).

Considerando a combinação *mt+sim* para os casos sem balanceamento, é possível observar que apenas um subconjunto – com $p = 99$ – obteve desempenho superior do que usar C_{all} como conjunto de treinamento, tanto para acurácia quanto para F_1. Nas situações em que o conjunto de treinamento é balanceado, nenhum subconjunto obteve resultados melhores de quando C_{all} é o conjunto de treinamento.

Analisando a configuração *mt+dis*, isto é, utilizando a média dos *embeddings* dos *tokens* (*mt*) para representação da base-alvo e selecionando as instâncias mais similares e as mais dissimilares (*dis*), para as situações sem balanceamento, é possível notar que quatro subconjuntos – quando $p = 99$, $p = 95$, $p = 96$, $p = 94$ – obtiveram desempenho melhor do usar C_{all} como treinamento, tanto para acurácia quanto para F_1. Nos casos com balanceamento, todos os cinco melhores subconjuntos – quando $p = 94$, $p = 98$, $p = 97$, $p = 95$, e $p = 93$ – apresentaram desempenhos superiores do que usar C_{all} como treinamento, em termos de acurácia, e nos subconjuntos quando $p = 98$, $p = 94$, $p = 97$, $p = 95$ e $p = 100$, em termos de F_1.

Tabela 3. Subconjuntos da base-fonte (%) com melhores desempenhos

Sem balanceamento					**Com balanceamento**				
Pos.	Acurácia		*F_1-measure*		Pos.	Acurácia		*F_1-measure*	
	% sel.	Ganho	% sel.	Ganho		% sel.	Ganho	% sel.	Ganho
Combinação *mt+sim*									
1º	**99%**	**1,0338**	**99%**	**1,1261**	1º	66%	1,0311	95%	1,1224
2º	100%	1,0332	100%	1,1253	2º	65%	1,0311	98%	1,1223
3º	96%	1,0329	96%	1,1249	3º	95%	1,0311	84%	1,1222
4º	97%	1,0325	98%	1,1243	4º	85%	1,0311	97%	1,1222
5º	98%	1,0324	97%	1,1243	5º	98%	1,0310	96%	1,1221
Combinação *mt+dis*									
1º	**99%**	**1,0340**	**99%**	**1,1265**	1º	**94%**	**1,0370**	**98%**	**1,1290**
2º	**95%**	**1,0336**	**95%**	**1,1258**	2º	**98%**	**1,0367**	**94%**	**1,1290**
3º	**96%**	**1,0336**	**96%**	**1,1258**	3º	**97%**	**1,0365**	**97%**	**1,1287**
4º	**94%**	**1,0334**	**94%**	**1,1257**	4º	**95%**	**1,0364**	**95%**	**1,1285**
5º	100%	1,0332	93%	1,1253	5º	**93%**	**1,0363**	**100%**	**1,1284**
Combinação *mi+sim*									
1º	**99%**	**1,0335**	**99%**	**1,1257**	1º	86%	1,0315	98%	1,1225
2º	100%	1,0332	100%	1,1253	2º	87%	1,0314	86%	1,1225
3º	93%	1,0332	93%	1,1250	3º	85%	1,0312	99%	1,1224
4º	92%	1,0330	96%	1,1249	4º	84%	1,0312	87%	1,1224
5º	96%	1,0329	98%	1,1247	5º	73%	1,0312	94%	1,1224
Combinação *mi+dis*									
1º	**93%**	**1,0339**	**93%**	**1,1266**	1º	**98%**	**1,0368**	**98%**	**1,1289**
2º	**91%**	**1,0336**	**91%**	**1,1262**	2º	**94%**	**1,0367**	**94%**	**1,1287**
3º	**92%**	**1,0336**	**92%**	**1,1262**	3º	**99%**	**1,0366**	**100%**	**1,1287**
4º	100%	1,0332	**89%**	**1,1255**	4º	**100%**	**1,0365**	**99%**	**1,1286**
5º	94%	1,0332	100%	1,1253	5º	**92%**	**1,0364**	**96%**	**1,1284**

Quanto à configuração *mi+sim*, apenas um subconjunto – com $p = 99$, sem balanceamento – apresenta ganho superior ao obtido com o treinamento realizado com C_{all}, tanto para acurácia quanto para F_1.

Por último, para a configuração *mi+dis*, analisando os casos sem balanceamento, três subconjuntos – com $p = 93$, $p = 91$ e $p = 92$ – apresentam resultados melhores do que usar C_{all} como treinamento, tanto para acurácia quanto para F_1. No entanto, em termos de F_1, o subconjunto formado por $p = 89$ das instâncias também apresentou desempenho superior. Para os casos com balanceamento, os cinco melhores subconjuntos – com $p = 98$, $p = 94$, $p = 99$, $p = 100$ e $p = 92$, em termos de acurácia e $p = 98$, $p = 94$, $p = 100$, $p = 99$ e $p = 96$ em termos de F_1 – apresentam desempenho superior ao obtido utilizando C_{all} como treinamento.

Na avaliação geral, considerando os maiores entre todos os valores de ganho

médio obtidos (valores sublinhados), é possível notar que a configuração *mt+dis com balanceamento* apresentou os melhores resultados. Especificamente, em termos de acurácia, quando $p = 94$, obteve-se ganho de 1,0370 e, em termos de F_1, com $p = 94$ e $p = 98$ foi obtido um ganho de 1,1290. No entanto, cabe destacar que os resultados reportados na Tabela 3, que correspondem aos subconjuntos da base-fonte com melhores desempenhos, foram obtidos com percentuais muito próximos a 100%, dando evidências de que essa estratégia de seleção de instâncias não foi efetiva.

4.2.2. Estratégia de Seleção S2

A estratégia S2 consiste na seleção das k instâncias da base-fonte mais similares a cada instância da base-alvo. Devido ao espaço limitado, são reportados apenas os melhores resultados, obtidos com o balanceamento da base-fonte, considerando que essa situação apresentou o melhor desempenho geral. A Tabela 4 apresenta os melhores resultados obtidos para cada base-alvo, variando o valor de k entre 1 e 20. Especificamente, para cada instância da base-alvo, são selecionadas as k instâncias da base-fonte mais similares a cada uma delas, tal que $1 \leq k \leq min(k_{max}, 20)$, em que k_{max} é definido a seguir.

O número de instâncias da base-fonte de cada classe a serem selecionadas, n_{sel}, para um determinado k, é dado por $n_{sel} = (k \times n_{alvo})/2$, em que n_{alvo} é o número de instâncias da base-alvo. Dessa forma, sabendo que $n_{sel} \leq n_{min}$, em que n_{min} é o número de instâncias que pertencem à classe minoritária na base-fonte, temos que $k \leq (2 \times n_{min})/n_{alvo}$. Logo, o valor máximo para k, k_{max}, é definido por $k_{max} = \lfloor (2 \times n_{min})/n_{alvo} \rfloor$.

Tabela 4. Valores de k com melhores desempenhos.

Bases de dados	k_{max}	Acurácia (ganho)		F_1-*measure* (ganho)	
		Melhor k x Alvo	21D x Alvo	Melhor k x Alvo	21D x Alvo
irony	20	20 (1,0692)	**1,1621**	20 (1,3528)	**1,4472**
sarcasm	20	2 (1,3479)	1,3479	10 (**1,7826**)	1,7814
aisopos	20	6 (**1,0890**)	1,0845	6 (**1,1083**)	1,1019
SemEval15-Fig	20	5 (**0,7081**)	0,6934	5 (**0,8424**)	0,8287
sentiment140	20	17 (1,0518)	**1,0691**	17 (1,0536)	**1,0709**
person	20	20 (1,0930)	**1,1347**	20 (1,3227)	**1,3549**
hobbit	20	20 (**1,1101**)	1,1073	20 (1,3457)	**1,3516**
iphone	20	14 (1,0295)	**1,0619**	14 (1,2727)	**1,3102**
movie	20	16 (0,9891)	**1,0065**	16 (1,1056)	**1,1228**
sanders	20	15 (**1,0247**)	1,0141	15 (**1,0338**)	1,0229
Narr	20	20 (1,0589)	**1,0716**	20 (1,0660)	**1,0774**
archeage	20	4 (**0,9737**)	0,9348	4 (**0,9806**)	0,9438
SemEval18	20	16 (**1,0528**)	1,0452	16 (**1,0605**)	1,0533
OMD	20	19 (**0,9454**)	0,9411	19 (**0,9506**)	0,9505
HCR	20	18 (**1,0372**)	1,0300	2 (**1,1669**)	1,1132
STS-gold	19	2 (**0,9899**)	0,9109	2 (**1,0492**)	0,9680
SentiStrength	17	13 (**1,0519**)	1,0478	13 (**1,0616**)	1,0564
Target-dependent	10	6 (**0,9746**)	0,9659	6 (**0,9747**)	0,9657
Vader	8	8 (1,0041)	**1,0121**	8 (1,0324)	**1,0399**
SemEval13	8	8 (1,0629)	**1,0781**	6 (1,1425)	**1,1527**
SemEval17	5	2 (**1,0028**)	0,9947	2 (**1,0079**)	1,0014
SemEval16	2	2 (1,0118)	**1,0177**	2 (1,0372)	**1,0424**

Na Tabela 4, a segunda coluna apresenta os valores máximos de k para cada base-alvo. A terceira e quinta colunas indicam os valores de k que produziram os melhores ganhos, em termos de acurácia e F_1, respectivamente, em relação ao desempenho obtido com o classificador treinado com a própria base-alvo. Entre parênteses são apresentados os valores desses ganhos. A quarta e sexta colunas apresentam os valores dos ganhos

obtidos ao usar toda a base-fonte para treinamento (21D), em termos de acurácia e F_1, respectivamente. Em negrito estão destacados os melhores valores de ganho.

Ao analisar os ganhos reportados na Tabela 4, é possível observar que a estratégia de seleção de instâncias S2 apresentou melhor desempenho em 12 das 22 bases (colunas *Melhor k x Alvo*), tanto para acurácia quanto para F_1, em relação ao desempenho utilizando toda a base-fonte (colunas *21D x Alvo*). Além disso, os valores de k que produzem os melhores resultados estão entre os mais próximos de k_{max}. Analisando as acurácias, para sete bases-alvo, o melhor valor de k é igual a k_{max} e para outras duas bases, OMD e HCR, o melhor valor de k se aproxima a k_{max} ($k \geq k_{max} - 2$). Em termos de F_1, para seis bases o melhor *k* é igual a k_{max} e para outras duas, OMD e SemEval13, o melhor desempenho também acontece para $k \geq k_{max} - 2$.

Assim como para os resultados obtidos pela estratégia de seleção S1 (Seção 4.2.1), é possível notar uma tendência de aumento no desempenho preditivo da estratégia S2 quanto maior é o conjunto de treinamento, dando evidências de que este tipo de seleção também não foi efetiva.

5. Conclusões e Trabalhos Futuros

Este artigo teve como objetivo determinar se vale a pena treinar um classificador de polaridade com a união de um conjunto de bases-fonte disponíveis, ou se é melhor selecionar uma base-fonte específica (Q1), como feito em [Guimarães et al. 2020], e para verificar se vale a pena selecionar um subconjunto de instâncias da união das bases-fonte disponíveis, com base na similaridade em relação às instâncias da base-alvo (Q2).

Para responder Q1, o experimento realizado apontou que usar todas as bases como fonte para o treinamento do classificador alcança, em geral, melhores resultados que selecionar uma única base, uma vez que o ganho médio obtido foi superior ao observado em [Guimarães et al. 2020].

Para responder Q2, duas formas de seleção de instâncias foram analisadas: selecionando um percentual das instâncias da base-fonte (estratégia S1) ou as k instâncias mais similares a cada instância da base-alvo (estratégia S2). Em relação à estratégia S1, foi possível notar que a seleção de instâncias da base-fonte para compor o conjunto de treinamento pode gerar classificadores com um ganho médio melhor do que o ganho obtido ao usar toda a base-fonte, porém com resultados muito próximos. No entanto, os percentuais que devem ser selecionados são, em sua maioria, elevados (próximos a 100%), não compensando o custo computacional de selecionar as instâncias. As configurações que obtiveram o melhor desempenho neste experimento foram as que selecionavam as instâncias mais similares e as mais dissimilares (*mt+dis* e *mi+dis*), e consideravam o balanceamento da base-fonte. Isso indica que a diversidade e o balanceamento do conjunto de treinamento podem influenciar positivamente no desempenho do classificador.

Quanto à estratégia S2, mais uma vez houve a indicação de que, ao aumentar o conjunto de treinamento, o desempenho do classificador tende a melhorar. Embora nesta avaliação existam situações em que o desempenho da seleção tenha apresentado melhores resultados do que usar toda a base-fonte, isso ocorre com uma pequena diferença de desempenho na maioria dos casos. Contudo, como para algumas bases o melhor desempenho foi obtido com $k = 20$, é possível obter um desempenho melhor ao aumentar o valor de k.

Em trabalhos futuros, podem ser exploradas outras métricas de distância para seleção de instâncias e outras formas de representações dos dados. Tendo em vista que incluir instâncias dissimilares mostrou-se promissor, novos trabalhos podem focar em ajustar a razão entre instâncias similares e dissimilares que estão sendo utilizadas.

Referências

Barbosa, L. and Feng, J. (2010). Robust sentiment detection on twitter from biased and noisy data. In *Proc. of the 23rd Int. Conf. on Computational Linguistics: Posters*, COLING '10, page 36–44. ACL.

Bravo-Marquez, F., Frank, E., Mohammad, S. M., and Pfahringer, B. (2016). Determining word-emotion associations from tweets by multi-label classification. In *Proc. of the 2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 536–539. IEEE.

Carvalho, J. and Plastino, A. (2021). On the evaluation and combination of state-of-the-art features in twitter sentiment analysis. *Artificial Intelligence Review*, 54.

Dong, L., Wei, F., Tan, C., Tang, D., Zhou, M., and Xu, K. (2014). Adaptive recursive neural network for target-dependent Twitter sentiment classification. In *Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 49–54. ACL.

Guimarães, E., Carvalho, J., Paes, A., and Plastino, A. (2020). Transfer learning for twitter sentiment analysis: Choosing an effective source dataset. In *Anais do VIII Symposium on Knowledge Discovery, Mining and Learning*, pages 161–168. SBC.

Guo, J., Shah, D., and Barzilay, R. (2018). Multi-source domain adaptation with mixture of experts. In *Proc. of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4694–4703. ACL.

Li, S., Song, S., and Huang, G. (2017). Prediction reweighting for domain adaptation. *IEEE Transactions on Neural Networks and Learning Systems*, 28(7):1682–1695.

Liu, B. (2020). *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Studies in Natural Language Processing. Cambridge University Press, 2 edition.

Martínez-Cámara, E., Martín-Valdivia, M., López, L., and Montejo-Ráez, A. (2014). Sentiment analysis in twitter. *Natural Language Engineering*, 20:1–28.

Mikolov, T., Chen, K., Corrado, G. S., and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *CoRR*, abs/1301.3781.

Pan, S. J. and Yang, Q. (2010). A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

Ruder, S. and Plank, B. (2017). Learning to select data for transfer learning with Bayesian Optimization. In *Proc. of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 372–382. ACL.

Santos, J. S., Paes, A., and Bernardini, F. (2019). Combining labeled datasets for sentiment analysis from different domains based on dataset similarity to predict electors sentiment. In *2019 8th Brazilian Conf. on Intelligent Systems*, pages 455–460.