Enriching datasets for sentiment analysis in tweets with instance selection

Eliseu Guimarães[1,2], Daniela Vianna[3], Aline Paes[1], Alexandre Plastino[1]

[1] Universidade Federal Fluminense, Brazil
[2] Marinha do Brasil
eliseuguimaraes@id.uff.br {alinepaes,plastino}@ic.uff.br
[3] Pesquisadora Independente
dvianna@gmail.com

Abstract. Sentiment analysis in tweets is a research field of great importance, mainly due to the popularity of Twitter. However, collecting and annotating tweets is an expensive and time-consuming task, making that some domains have only a limited set of labeled data. A promising strategy to handle this issue is to leverage labeled domains rich in data to select instances that enrich target datasets. This paper proposes different strategies for selecting instances from a set of labeled source datasets in order to improve the performance of classifiers trained only with the target dataset. Different approaches are proposed, including similarity metrics and variations in the number of selected instances. The results show that the size of the training set plays an essential role in the predictive capacity of the classifier. Furthermore, the results point out the importance of taking into account diversity criteria when selecting the instances.

CCS Concepts: • **Computing methodologies** → **Transfer learning**.

Keywords: machine learning, sentiment analysis, supervised learning, transfer learning

1. INTRODUÇÃO

A análise de sentimentos é o estudo computacional das opiniões, sentimentos, emoções e atitudes das pessoas [Liu 2020]. Com a crescente popularização das redes sociais, esse campo de pesquisa tem se tornado cada vez mais importante, visto que as pessoas são incentivadas a emitirem opiniões sobre os mais diversos assuntos. Uma dessas redes, o Twitter[1], um serviço de microblog de textos curtos, chamados tweets, apresenta desafios próprios, como a presença de linguagem informal, a utilização de palavras grafadas de forma incorreta e a falta de contexto [Martínez-Cámara et al. 2014].

Uma das tarefas que a análise de sentimentos abrange é a detecção da polaridade de opiniões. No caso específico deste estudo, é tratada a detecção de polaridade em tweets. Abordagens baseadas em aprendizado de máquina são vastamente usadas para tratar essa tarefa, extraindo características dos tweets e as utilizando como atributos para o treinamento de classificadores. Em geral, dados de um determinado domínio são utilizados para treinar classificadores para o mesmo domínio. Contudo, há situações em que os dados rotulados disponíveis em um domínio não são suficientes para treinar um classificador com bom desempenho, seja devido ao fato de o domínio de interesse ser raro, ou por ser proibitivo rotular manualmente os dados existentes, ou ainda porque falta qualidade aos dados.

Para lidar com esse problema, uma abordagem oriunda da área de transferência de aprendizado [Pan and Yang 2010] é selecionar instâncias a partir de domínios-fonte para enriquecer o conjunto de

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

treinamento do classificador associado ao domínio-alvo, de modo a aumentar o seu desempenho preditivo [Guo et al. 2018; Liu et al. 2019; Ruder et al. 2017; Ruder and Plank 2017]. Porém, a maioria dos trabalhos anteriores requerem treinamento de métricas ou divisão das bases-fonte em subconjuntos. Além disso, os trabalhos anteriores não lidam especificamente com análise de sentimentos em tweets.

Este artigo investiga três abordagens de seleção de dados de um conjunto de bases-fonte oriundas de diversos domínios, com o objetivo de enriquecer o conjunto de treinamento para detecção de polaridade em uma base-alvo. No cenário investigado aqui, a base-alvo possui tweets rotulados para o treinamento de um classificador mas deseja-se melhorar o seu desempenho preditivo com um conjunto de dados enriquecido. Os resultados dos experimentos mostram que utilizar instâncias selecionadas de um conjunto de bases-fonte para enriquecer o conjunto de treinamento produz um aumento no desempenho dos classificadores em comparação com o treinamento apenas com a base-alvo. Esse aumento ocorre especialmente quando a seleção é composta de uma combinação de instâncias mais próximas e de instâncias mais distantes de cada instância da base-alvo.

O restante deste artigo se estrutura como segue. Na Seção 2, são apresentados trabalhos relacionados ao estudo desenvolvido, enquanto na Seção 3 é mostrada a metodologia utilizada por esta pesquisa. Na Seção 4, os resultados dos experimentos são apresentados, com suas respectivas conclusões sendo debatidas na Seção 5, onde ainda são apontados trabalhos futuros.

2. TRABALHOS RELACIONADOS

Diversos trabalhos têm proposto abordagens distintas para resolver o problema de seleção de dados de treinamento a partir de uma ou mais bases-fonte com o objetivo de treinar classificadores mais robustos para uma base-alvo de domínio distinto [Guo et al. 2018; Liu et al. 2019; Ruder et al. 2017; Ruder and Plank 2017]. Em [Guo et al. 2018], é utilizada uma abordagem do tipo *mixture-of-experts*, com diversas bases-fonte. Nesse trabalho, é considerado que cada base-fonte está alinhada a uma região distinta da base-alvo e uma métrica *point-to-set* é aprendida para ponderar os resultados de classificadores treinados com essas bases-fonte. O estudo conclui que as acurácias obtidas com esse tipo de estratégia foram superiores a se utilizar apenas uma base-fonte ou a união de todas as bases-fonte.

Por sua vez, [Liu et al. 2019] propõe uma abordagem de aprendizado por reforço, em que um *framework*, formado por dois componentes, busca instâncias relevantes e aprende melhores representações para elas. Um dos componentes é responsável por selecionar dados considerando um vetor de distribuição baseado na seleção de dados do passo anterior, enquanto o outro é responsável por fazer a extração de atributos dos dados, atualizar as recompensas para a geração do vetor de distribuição e gerar o classificador para a tarefa. Os resultados mostraram que esta abordagem teve um melhor desempenho em três de quatro bases-alvo utilizadas, em comparação com outros estudos.

Em [Ruder et al. 2017], são analisadas estratégias de seleção de dados, onde se consideram três fatores importantes para a seleção: a representação dos dados, a métrica de similaridade e o nível de seleção. Para cada uma das três representações avaliadas, é utilizada a métrica de similaridade mais comumente associada a ela. Os resultados apontam que utilizar seleção de subconjuntos de instâncias pode ter melhor desempenho preditivo do que a seleção individual de instâncias.

A abordagem proposta em [Ruder and Plank 2017] utiliza otimização Bayesiana para aprender uma métrica de similaridade de bases assumindo que diferentes tarefas e diferentes domínios pressupõem diferentes noções de similaridade. Foram utilizadas seis métricas de similaridade entre bases, três tipos de representações de dados e seis métricas de diversidade aplicadas ao conjunto de treinamento. O trabalho conclui que utilizar métricas de diversidade junto com métricas de similaridade melhora o desempenho preditivo, superando os resultados que selecionam aleatoriamente dados ou usam uma única métrica.

Em comparação à literatura, neste estudo são apresentadas estratégias de seleção de dados que

não requerem treinamento de métricas ou divisão das bases-fonte em subconjuntos. Adicionalmente, trata-se de um estudo específico de análise de sentimentos em tweets. Destaca-se ainda o fato de ser utilizado um grande e diverso conjunto de bases de diferentes domínios, o que confere robustez aos resultados.

3. METODOLOGIA

Nesta seção, está descrita a metodologia utilizada. Na Subseção 3.1, as bases utilizadas são descritas e são apresentados os procedimentos de pré-processamento executados para a extração dos atributos. A Subseção 3.2 descreve os procedimentos adotados na condução dos experimentos.

3.1 Bases de dados e pré-processamento

Nas avaliações conduzidas, utiliza-se um conjunto de 22 bases de dados de tweets em língua inglesa[2] [Carvalho and Plastino 2021]. As características dessas bases são apresentadas na Tabela I. Como pré-processamento dos tweets, inicialmente as menções a usuários e as URLs foram substituídas por expressões únicas. Os tweets foram, em seguida, tokenizados e colocados em letras minúsculas. Os atributos foram obtidos utilizando *word embeddings*, a partir de um modelo estático [Bravo-Marquez et al. 2016] que possui bom desempenho para análise de sentimentos em tweets [Carvalho and Plastino 2021]. O cálculo dos atributos de cada instância foi realizado computando a média dos *embeddings* referentes aos tokens da instância e, caso algum token não tivesse correspondência no modelo pré-treinado, seus *embeddings* foram considerados como um vetor nulo.

Base	Abreviação	#pos	#neg	% pos	Total	Base	Abreviação	#pos	#neg	% pos	Total
irony	iro	22	43	34%	65	archeage	arc	724	994	42%	1718
sarcasm	sar	33	38	46%	71	SemEval18	S18	865	994	47%	1859
aisopos	ais	159	119	57%	278	OMD	OMD	710	1196	37%	1906
SemEval15-Fig	S15	47	274	15%	321	HCR	HCR	539	1369	28%	1908
sentiment140	sem	182	177	51%	359	STS-gold	STS	632	1402	31%	2034
person	per	312	127	71%	439	SentiStrength	SSt	1340	949	59%	2289
hobbit	hob	354	168	68%	522	Target-dependent	Tar	1734	1733	50%	3467
iphone	iph	371	161	70%	532	Vader	Vad	2897	1299	69%	4196
movie	mov	460	101	82%	561	SemEval13	S13	3183	1195	73%	4378
sanders	san	570	654	47%	1224	SemEval17	S17	2375	3972	37%	6347
Narr	nar	739	488	60%	1227	SemEval16	S16	8893	3323	73%	12216

Tabela I. Características das bases de dados.

3.2 Procedimentos experimentais

Nos experimentos realizados, foi utilizado o algoritmo SVM (Support Vector Machines), na sua implementação do scikit-learn [Pedregosa et al. 2011], com o parâmetro de ponderação de classe configurado para a forma balanceada devido a seu bom desempenho em análise de sentimentos em tweets [Barreto et al. 2021]. Além disso, foram considerados como *baselines* os valores de acurácia e de F_1 ponderados obtidos a partir de um procedimento de validação cruzada estratificada com 10 *folds* utilizando como conjunto de treinamento a própria base-alvo. Cada experimento foi realizado considerando cada uma das bases como base-alvo e as 21 restantes como a união das bases-fonte.

O primeiro experimento visava verificar se enriquecer o conjunto de treinamento com a maior quantidade possível de instâncias da união de bases-fonte, garantindo o balanceamento, produz melhora de desempenho em comparação com os *baselines*. Neste experimento, a base-alvo foi dividida nas mesmas 10 partições utilizadas para a geração dos valores *baseline*. Cada partição foi separada como teste e as nove restantes foram usadas como parte do conjunto de treinamento, que foi completado aleatoriamente com instâncias da união das bases-fonte. O modelo foi aplicado à partição de teste e

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

o procedimento foi repetido para todas as partições, sendo calculadas a média da acurácia e do F_1 ponderado para se obter o desempenho final. A comparação com o *baseline* se dá pelo cálculo do ganho, computado como a razão entre o valor da métrica de desempenho obtida quando o conjunto de treinamento é formado pela base-alvo enriquecida pelas instâncias da união das bases-fonte e o valor do *baseline*.

O segundo experimento também considera a união das bases-fonte na construção do conjunto de treinamento. Entretanto, nesse experimento, diferentes métricas de seleção de instâncias são investigadas. O objetivo é identificar se existe um subconjunto da união das bases-fonte que, quando agregadas ao conjunto de treinamento, produz um classificador com poder preditivo superior ao *baseline*. Para este experimento, a base-alvo também foi dividida nas mesmas 10 partições utilizadas para a geração do *baseline*, sendo adotado um procedimento semelhante ao do experimento anterior. Porém, neste experimento, inicialmente é calculada a quantidade de instâncias que precisa ser agregada à classe minoritária das partições de treinamento para que o conjunto de treinamento fique balanceado. Considerando-se que esta quantidade vai ser retirada da união das bases-fonte, verifica-se qual será a classe minoritária do restante da união das bases-fonte e calcula-se a quantidade de instâncias dessa classe que deve ser agregada ao conjunto de treinamento, de forma a atender a um percentual de seleção.

São, então, adicionadas instâncias da união de bases-fonte segundo três critérios: (I) seleção aleatória de instâncias, (II) seleção das instâncias mais próximas a cada instância das partições que formam o conjunto de treinamento, (III) seleção das instâncias mais próximas e mais distantes a cada instância das partições que formam o conjunto de treinamento. Neste último critério, foram selecionadas quantidades iguais de instâncias mais próximas e instâncias mais distantes. Para os critérios (II) e (III), foi adotada como métrica de similaridade a distância Euclidiana. Ainda para estes dois últimos critérios, para cada instância das partições de treinamento da base-alvo só eram selecionadas instâncias da união da base-fonte que tivessem a mesma classe da instância da base-alvo. Para os três critérios, o balanceamento de classes foi sempre mantido. Os modelos gerados foram aplicados à partição de teste e seus resultados comparados com os *baselines*.

4. RESULTADOS

Nesta seção, são apresentados os resultados obtidos com os experimentos descritos na Seção 3. A Tabela II apresenta os resultados do primeiro experimento. Nela, são mostrados as acurácias e os F_1 ponderados obtidos quando o conjunto de treinamento é formado apenas pela base-alvo (colunas Ac_a e F_{1-a}) e quando ele é formado pela base-alvo em conjunto com a união das bases-fonte (colunas Ac_{a+f} e F_{1-a+f}). As colunas "Ganho Ac." e "Ganho F_1" apresentam os ganhos de acurácia e F_1, isto é, os resultados das divisões das colunas Ac_{a+f} e F_{1-a+f} pelas colunas Ac_a e F_{1-a}, respectivamente. Estão assinalados em negrito os valores de ganho maiores ou iguais a 1, ou seja, aqueles valores que indicam que o desempenho utilizando a união das bases-fonte em conjunto com a base-alvo superou ou igualou o uso apenas da base-alvo. Tanto para a acurácia quanto para F_1 foram 15 as bases para as quais isso ocorreu. Cabe ressaltar ainda que, para a maioria das bases, os ganhos (razões) foram muito próximos a 1, o que significa que a diferença de desempenho não foi significativa.

As Tabelas III-VI apresentam os resultados do segundo experimento. Na Tabela III, são apresentados os valores de acurácia e F_1 obtidos quando o conjunto de treinamento é formado por instâncias da base-alvo e instâncias da união das bases-fonte selecionadas de forma aleatória. O tamanho dessas seleções é definido percentualmente e, conforme pode ser observado na tabela, varia de 0,0% a 100,0%, sendo este último percentual o equivalente ao experimento anterior.

A Tabela IV apresenta os ganhos obtidos considerando a seleção feita de forma aleatória, ou seja, os valores mostrados nesta tabela são os valores da Tabela III divididos pelos seus respectivos *baselines*, estando assinalados em negrito os casos em que o ganho é maior ou igual a 1. Nas três últimas

Base	Ac_a	Ac_{a+f}	Ganho Ac.	F_{1-a}	F_{1-a+f}	Ganho F_1	Base	Ac_a	Ac_{a+f}	Ganho Ac.	F_{1-a}	F_{1-a+f}	Ganho F_1
iro	0,63	0,77	**1,22**	0,62	0,74	**1,19**	sar	0,69	0,85	**1,22**	0,67	0,85	**1,27**
ais	0,94	0,95	**1,00**	0,94	0,95	**1,00**	S15	0,90	0,76	0,84	0,90	0,78	0,87
sem	0,87	0,87	**1,01**	0,87	0,87	**1,01**	per	0,78	0,82	**1,06**	0,79	0,82	**1,05**
hob	0,89	0,83	0,93	0,89	0,83	0,93	iph	0,79	0,78	0,98	0,80	0,79	0,99
mov	0,83	0,86	**1,05**	0,83	0,87	**1,04**	san	0,83	0,84	**1,00**	0,83	0,84	**1,00**
nar	0,88	0,91	**1,04**	0,88	0,91	**1,04**	arc	0,87	0,85	0,99	0,87	0,85	0,98
S18	0,83	0,83	**1,00**	0,83	0,83	**1,00**	OMD	0,84	0,81	0,96	0,84	0,80	0,96
HCR	0,75	0,78	**1,04**	0,76	0,75	0,99	STS	0,86	0,86	0,99	0,86	0,86	**1,00**
SSt	0,80	0,82	**1,02**	0,80	0,82	**1,02**	Tar	0,83	0,82	0,98	0,83	0,82	0,98
Vad	0,87	0,88	**1,00**	0,88	0,88	**1,00**	S13	0,81	0,84	**1,05**	0,81	0,84	**1,03**
S17	0,88	0,88	**1,01**	0,88	0,88	**1,00**	S16	0,85	0,86	**1,02**	0,85	0,86	**1,02**

Tabela II. Acurácias e F_1 obtidos com a base-alvo (a) e base-alvo+base-fonte (a+f), com seus respectivos ganhos.

Base	Acurácia										F_1									
	Percentuais selecionados										Percentuais selecionados									
	0,0	0,5	1,0	2,5	5,0	10,0	20,0	40,0	80,0	100,0	0,0	0,5	1,0	2,5	5,0	10,0	20,0	40,0	80,0	100,0
iro	0,65	0,66	0,66	0,72	0,77	0,74	0,75	0,75	0,77	0,77	0,64	0,64	0,64	0,71	0,74	0,71	0,73	0,73	0,74	0,74
sar	0,69	0,79	0,76	0,75	0,75	0,82	0,78	0,86	0,85	0,85	0,67	0,78	0,75	0,74	0,73	0,81	0,77	0,85	0,84	0,85
ais	0,93	0,92	0,92	0,91	0,92	0,92	0,94	0,94	0,94	0,95	0,93	0,92	0,92	0,91	0,92	0,92	0,94	0,94	0,94	0,95
S15	0,88	0,88	0,88	0,87	0,87	0,85	0,83	0,79	0,76	0,76	0,89	0,88	0,88	0,87	0,87	0,85	0,84	0,80	0,78	0,78
sem	0,87	0,87	0,87	0,87	0,88	0,88	0,89	0,87	0,87	0,87	0,87	0,87	0,87	0,87	0,88	0,88	0,89	0,87	0,87	0,87
per	0,77	0,79	0,80	0,80	0,82	0,81	0,81	0,82	0,82	0,82	0,77	0,79	0,80	0,80	0,82	0,81	0,81	0,82	0,82	0,82
hob	0,88	0,88	0,88	0,86	0,87	0,84	0,83	0,83	0,83	0,83	0,88	0,88	0,88	0,86	0,86	0,84	0,83	0,83	0,82	0,83
iph	0,80	0,80	0,81	0,81	0,79	0,78	0,79	0,78	0,78	0,78	0,80	0,80	0,82	0,82	0,80	0,79	0,80	0,79	0,79	0,79
mov	0,84	0,87	0,86	0,85	0,84	0,86	0,85	0,86	0,86	0,86	0,83	0,86	0,85	0,85	0,84	0,86	0,85	0,86	0,86	0,87
san	0,84	0,84	0,84	0,84	0,83	0,83	0,83	0,83	0,83	0,84	0,84	0,84	0,84	0,84	0,83	0,83	0,83	0,83	0,83	0,84
nar	0,88	0,89	0,89	0,89	0,89	0,90	0,90	0,91	0,91	0,91	0,88	0,89	0,89	0,89	0,90	0,90	0,90	0,91	0,91	0,91
arc	0,87	0,87	0,87	0,87	0,87	0,86	0,85	0,86	0,86	0,85	0,87	0,87	0,87	0,87	0,86	0,86	0,85	0,86	0,86	0,85
S18	0,83	0,83	0,83	0,84	0,84	0,83	0,83	0,83	0,83	0,83	0,83	0,83	0,83	0,84	0,84	0,83	0,83	0,83	0,83	0,83
OMD	0,83	0,83	0,83	0,83	0,83	0,84	0,82	0,82	0,81	0,81	0,83	0,82	0,83	0,83	0,83	0,83	0,82	0,81	0,81	0,80
HCR	0,79	0,79	0,79	0,80	0,80	0,79	0,79	0,79	0,79	0,78	0,77	0,77	0,77	0,77	0,77	0,76	0,76	0,76	0,76	0,75
STS	0,86	0,87	0,87	0,87	0,87	0,87	0,87	0,87	0,86	0,86	0,86	0,87	0,87	0,87	0,87	0,87	0,87	0,87	0,86	0,86
SSt	0,81	0,81	0,81	0,81	0,81	0,81	0,82	0,81	0,82	0,82	0,81	0,81	0,81	0,81	0,81	0,81	0,82	0,81	0,82	0,82
Tar	0,83	0,83	0,83	0,83	0,83	0,83	0,82	0,82	0,82	0,82	0,83	0,83	0,83	0,83	0,83	0,83	0,82	0,82	0,82	0,82
Vad	0,87	0,87	0,88	0,88	0,88	0,88	0,88	0,87	0,88	0,88	0,88	0,88	0,88	0,88	0,88	0,88	0,88	0,88	0,88	0,88
S13	0,82	0,83	0,83	0,83	0,84	0,84	0,84	0,84	0,85	0,84	0,83	0,83	0,83	0,83	0,84	0,84	0,84	0,84	0,84	0,84
S17	0,88	0,88	0,88	0,88	0,88	0,88	0,89	0,89	0,89	0,88	0,88	0,88	0,88	0,88	0,88	0,88	0,89	0,88	0,88	0,88
S16	0,85	0,86	0,86	0,86	0,86	0,86	0,86	0,86	0,86	0,86	0,85	0,86	0,86	0,86	0,86	0,86	0,86	0,86	0,86	0,86

Tabela III. Acurácias e F_1 obtidos com seleção aleatória de percentuais da base-fonte associados à base-alvo.

Base	Acurácia										F_1									
	Percentuais selecionados										Percentuais selecionados									
	0,0	0,5	1,0	2,5	5,0	10,0	20,0	40,0	80,0	100,0	0,0	0,5	1,0	2,5	5,0	10,0	20,0	40,0	80,0	100,0
iro	**1,03**	**1,06**	**1,05**	**1,15**	**1,22**	**1,17**	**1,20**	**1,20**	**1,22**	**1,22**	**1,02**	**1,03**	**1,03**	**1,13**	**1,18**	**1,13**	**1,17**	**1,17**	**1,19**	**1,19**
sar	**1,00**	**1,14**	**1,10**	**1,08**	**1,08**	**1,18**	**1,12**	**1,24**	**1,22**	**1,22**	**1,01**	**1,18**	**1,13**	**1,11**	**1,10**	**1,22**	**1,15**	**1,28**	**1,27**	**1,27**
ais	0,98	0,98	0,98	0,96	0,97	0,98	**1,00**	0,99	**1,00**	**1,00**	0,99	0,98	0,98	0,96	0,97	0,98	**1,00**	0,99	**1,00**	**1,00**
S15	0,98	0,98	0,98	0,96	0,97	0,94	0,92	0,87	0,84	0,84	0,99	0,98	0,98	0,96	0,97	0,95	0,93	0,89	0,87	0,87
sem	**1,00**	**1,00**	**1,00**	**1,01**	**1,02**	**1,01**	**1,03**	**1,00**	**1,00**	**1,01**	**1,00**	**1,00**	**1,00**	**1,01**	**1,02**	**1,01**	**1,03**	**1,00**	**1,00**	**1,01**
per	0,99	**1,02**	**1,03**	**1,03**	**1,06**	**1,04**	**1,04**	**1,06**	**1,06**	**1,06**	0,98	**1,00**	**1,02**	**1,02**	**1,04**	**1,03**	**1,03**	**1,05**	**1,04**	**1,05**
hob	0,98	0,99	0,98	0,97	0,97	0,95	0,94	0,94	0,93	0,93	0,99	0,99	0,98	0,97	0,97	0,94	0,93	0,93	0,92	0,93
iph	**1,01**	**1,01**	**1,03**	**1,03**	**1,00**	0,99	**1,00**	0,99	0,99	0,98	**1,00**	**1,01**	**1,03**	**1,03**	**1,00**	0,99	**1,00**	0,99	0,99	0,99
mov	**1,02**	**1,05**	**1,05**	**1,03**	**1,02**	**1,04**	**1,03**	**1,05**	**1,04**	**1,05**	**1,00**	**1,03**	**1,02**	**1,02**	**1,01**	**1,03**	**1,02**	**1,04**	**1,04**	**1,04**
san	**1,01**	**1,01**	**1,01**	**1,01**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,01**	**1,01**	**1,01**	**1,01**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**
nar	**1,00**	**1,01**	**1,02**	**1,02**	**1,02**	**1,03**	**1,03**	**1,04**	**1,04**	**1,04**	**1,00**	**1,01**	**1,02**	**1,02**	**1,02**	**1,03**	**1,03**	**1,04**	**1,04**	**1,04**
arc	**1,01**	**1,01**	**1,01**	**1,00**	**1,00**	**1,00**	0,98	0,99	0,99	0,99	**1,01**	**1,01**	**1,01**	**1,00**	**1,00**	**1,00**	0,98	0,99	0,99	0,98
S18	**1,00**	**1,00**	**1,00**	**1,00**	**1,01**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,01**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**
OMD	0,99	0,99	0,99	0,99	**1,00**	**1,00**	0,98	0,98	0,97	0,96	0,99	0,98	0,99	0,99	0,99	0,99	0,98	0,97	0,97	0,96
HCR	**1,04**	**1,05**	**1,05**	**1,06**	**1,06**	**1,05**	**1,05**	**1,05**	**1,04**	**1,04**	**1,01**	**1,01**	**1,01**	**1,02**	**1,01**	**1,01**	**1,00**	**1,00**	0,99	0,99
STS	0,99	**1,01**	**1,01**	**1,01**	**1,01**	**1,00**	**1,01**	**1,01**	0,99	0,99	0,99	**1,01**	**1,01**	**1,01**	**1,01**	**1,00**	**1,01**	**1,01**	**1,00**	**1,00**
SSt	**1,01**	**1,01**	**1,01**	**1,01**	**1,01**	**1,01**	**1,02**	**1,01**	**1,02**	**1,02**	**1,01**	**1,01**	**1,01**	**1,01**	**1,01**	**1,01**	**1,02**	**1,01**	**1,02**	**1,02**
Tar	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	0,99	0,99	0,98	0,98	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	0,99	0,99	0,98	0,98
Vad	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**
S13	**1,02**	**1,02**	**1,03**	**1,03**	**1,04**	**1,04**	**1,05**	**1,05**	**1,05**	**1,05**	**1,01**	**1,02**	**1,02**	**1,02**	**1,03**	**1,03**	**1,03**	**1,04**	**1,04**	**1,03**
S17	**1,00**	**1,00**	**1,00**	**1,00**	**1,01**	**1,01**	**1,01**	**1,01**	**1,01**	**1,01**	**1,00**	**1,00**	**1,00**	**1,00**	**1,01**	**1,01**	**1,01**	**1,01**	**1,01**	**1,00**
S16	**1,01**	**1,01**	**1,01**	**1,01**	**1,01**	**1,02**	**1,02**	**1,02**	**1,02**	**1,02**	**1,01**	**1,01**	**1,01**	**1,01**	**1,01**	**1,01**	**1,01**	**1,01**	**1,02**	**1,02**
									#Ganhos		16	**18**	**18**	**18**	**18**	17	17	15	15	15
									#Melhores		1	2	1	4	1	1	3	1	4	**5**
									Ranking		7,18	5,50	5,45	5,00	**4,64**	5,09	5,23	5,50	5,36	5,59

Tabela IV. Ganhos de acurácia e F_1 obtidos com seleção aleatória de percentuais da base-fonte associados à base-alvo.

linhas da tabela, são colocados, para cada percentual selecionado, as quantidades de ganhos de F_1 que são maiores ou iguais a 1 (#Ganhos), a quantidade de vezes que o percentual produziu o melhor desempenho em termos de F_1 dentre todos os percentuais analisados (#Melhores) e o ranking do percentual. Considerando apenas os valores de #Ganhos, a melhor seleção ocorreu com 0,5%, 1,0%,

2,5% e 5,0% da união das bases-fonte (18 ganhos), seguidas pela seleção de 10,0% e 20,0% (17 ganhos). No que diz respeito ao desempenho dos melhores ganhos (#Melhores), o melhor percentual foi 100,0% (5 melhores), seguido por 2,5% e 80,0% (4 melhores). Selecionar 5,0% da união das bases-fonte teve a melhor posição média no ranking (posição média de 4,64). Como nenhum percentual se mostrou claramente melhor, pode ser considerado que 2,5%, 5,0% e 100,0% tiveram o melhor desempenho geral para essa estratégia.

A Tabela V apresenta os resultados dos ganhos de acurácia e F_1 para a estratégia que seleciona as instâncias da união das bases-fonte pelo critério da proximidade por distância Euclidiana a cada uma das instâncias do conjunto de treinamento da base-alvo. Por questões de limitação de espaço, não será apresentada a tabela com os valores absolutos de acurácia e F_1. Com esta abordagem, as maiores quantidades de ganhos ocorreram com a seleção de 10,0% (19 ganhos) e 5,0% (18 ganhos) e, considerando o critério de melhores resultados, o percentual selecionado com melhor desempenho foi 20,0% (6 melhores). Este percentual também obteve a melhor posição média no ranking (3,23), e podemos considerar que a seleção de 10,0% ou 20,0% são as melhores para esta estratégia.

Base	Acurácia										F_1									
	Percentuais selecionados										Percentuais selecionados									
	0,0	0,5	1,0	2,5	5,0	10,0	20,0	40,0	80,0	100,0	0,0	0,5	1,0	2,5	5,0	10,0	20,0	40,0	80,0	100,0
iro	**1,15**	**1,20**	**1,18**	**1,18**	**1,15**	**1,20**	**1,23**	**1,20**	**1,25**	**1,33**	**1,15**	**1,19**	**1,15**	**1,16**	**1,13**	**1,18**	**1,22**	**1,17**	**1,23**	**1,32**
sar	**1,00**	**1,06**	**1,08**	**1,12**	**1,12**	**1,12**	**1,18**	**1,16**	**1,18**	**1,22**	**1,01**	**1,08**	**1,11**	**1,16**	**1,16**	**1,16**	**1,23**	**1,21**	**1,21**	**1,26**
ais	0,98	0,98	0,98	0,98	0,98	**1,00**	0,99	0,97	0,97	0,98	0,98	0,99	0,98	0,98	0,98	**1,00**	0,99	0,97	0,97	0,98
S15	0,96	0,97	0,96	0,94	0,94	0,92	0,90	0,89	0,84	0,81	0,97	0,98	0,96	0,95	0,95	0,93	0,92	0,91	0,87	0,85
sem	**1,01**	**1,01**	**1,00**	**1,01**	**1,01**	**1,01**	**1,02**	**1,02**	0,99	0,98	**1,01**	**1,01**	**1,00**	**1,01**	**1,01**	**1,01**	**1,02**	**1,02**	0,99	0,98
per	**1,01**	**1,01**	**1,01**	**1,03**	**1,03**	**1,07**	**1,05**	**1,04**	**1,06**	**1,07**	**1,00**	**1,00**	**1,00**	**1,02**	**1,03**	**1,06**	**1,04**	**1,03**	**1,05**	**1,05**
hob	0,99	0,99	0,98	0,97	0,97	0,96	0,96	0,95	0,92	0,92	0,99	0,99	0,98	0,97	0,97	0,95	0,96	0,94	0,92	0,91
iph	0,99	**1,01**	**1,01**	**1,01**	**1,03**	**1,03**	**1,04**	**1,03**	0,99	0,98	0,99	**1,01**	**1,01**	**1,01**	**1,03**	**1,02**	**1,04**	**1,03**	0,99	0,98
mov	**1,00**	**1,01**	**1,02**	**1,03**	**1,03**	**1,05**	**1,05**	**1,04**	**1,05**	**1,05**	**1,00**	**1,00**	**1,01**	**1,01**	**1,02**	**1,04**	**1,04**	**1,04**	**1,04**	**1,04**
san	**1,01**	**1,00**	**1,01**	**1,00**	**1,01**	**1,00**	**1,01**	**1,00**	0,99	0,97	**1,01**	**1,00**	**1,01**	**1,01**	**1,01**	**1,00**	**1,01**	**1,00**	0,99	0,97
nar	**1,00**	**1,00**	**1,00**	**1,01**	**1,01**	**1,02**	**1,03**	**1,03**	**1,01**	**1,02**	**1,00**	**1,00**	**1,00**	**1,01**	**1,01**	**1,02**	**1,03**	**1,02**	**1,01**	**1,02**
arc	**1,00**	**1,00**	0,99	0,99	**1,00**	**1,00**	0,99	0,99	0,97	0,97	**1,00**	**1,00**	0,99	0,99	**1,00**	**1,00**	0,99	0,99	0,97	0,97
S18	**1,00**	**1,01**	**1,00**	**1,01**	**1,00**	**1,00**	**1,01**	**1,01**	**1,00**	**1,01**	**1,00**	**1,01**	**1,00**	**1,01**	**1,00**	**1,00**	**1,01**	**1,01**	**1,00**	**1,01**
OMD	0,97	0,97	0,98	0,99	0,99	0,98	0,98	0,97	0,96	0,95	0,97	0,97	0,98	0,99	0,99	0,98	0,98	0,96	0,95	0,94
HCR	**1,02**	**1,03**	**1,02**	**1,04**	**1,04**	**1,04**	**1,04**	**1,05**	**1,04**	**1,04**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**
STS	0,99	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,01**	0,99	0,98	0,97	0,99	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,01**	0,99	0,98	0,97
SSt	**1,00**	**1,00**	**1,00**	**1,01**	**1,02**	**1,02**	**1,02**	**1,02**	**1,01**	**1,01**	**1,00**	**1,00**	**1,00**	**1,01**	**1,02**	**1,02**	**1,02**	**1,01**	**1,01**	**1,01**
Tar	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	0,98	0,98	0,98	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	0,98	0,98	0,98
Vad	0,99	0,99	0,99	0,99	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	0,99	0,99	0,99	0,99	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**
S13	**1,01**	**1,01**	**1,02**	**1,02**	**1,02**	**1,02**	**1,03**	**1,03**	**1,04**	**1,04**	**1,01**	**1,01**	**1,01**	**1,01**	**1,02**	**1,02**	**1,02**	**1,02**	**1,02**	**1,02**
S17	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,01**	**1,01**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,01**	**1,01**	**1,00**	**1,00**
S16	**1,01**	**1,01**	**1,01**	**1,01**	**1,01**	**1,01**	**1,02**	**1,02**	**1,02**	**1,02**	**1,00**	**1,00**	**1,00**	**1,01**	**1,01**	**1,01**	**1,01**	**1,01**	**1,02**	**1,01**
									#Ganhos		15	17	16	16	18	**19**	17	15	12	12
									#Melhores		2	1	0	2	1	3	**6**	3	2	2
									Ranking		7,36	6,05	6,73	5,05	4,64	3,95	**3,23**	4,77	6,41	6,50

Tabela V. Ganhos de acurácia e F_1 obtidos com seleção percentual das instâncias da base-fonte mais próximas à base-alvo.

Na Tabela VI, são apresentados os ganhos de acurácia e F_1 obtidos com a estratégia de selecionar as instâncias da união das bases-fonte que sejam mais próximas e as mais distantes a cada instância do conjunto de treinamento da base-alvo pelo critério da distância Euclidiana. Para esta estratégia, os melhores resultados para o critério de #Ganhos foram obtidos selecionando 1,0% e 2,5% (19 ganhos) e 10,0% e 20,0% (18 ganhos). Para #Melhores os melhores resultados foram produzidos selecionando 40,0% e 100,0% (5 melhores), o que explica o fato de 40,0% ter obtido a melhor posição média no ranking (3,86), sendo considerado o melhor percentual para esta estratégia. Um detalhe importante a considerar é que, como são selecionadas por esta estratégia quantidades iguais de instâncias mais próximas e mais distantes, selecionar 40,0% da união das bases-fonte é simplesmente acrescentar as 20,0% instâncias mais distantes na união das bases-fonte às 20,0% mais próximas que representam um dos melhores resultados da estratégia anterior.

Na Tabela VII, estão comparados os melhores resultados para cada uma das estratégias. Estão assinalados em negrito, novamente, valores de ganho maiores ou iguais a 1. As três últimas linhas são geradas levando em consideração apenas as seis combinações de estratégia-percentual colocadas nesta tabela. No que diz respeito ao número de ganhos maiores ou iguais a 1, a seleção das instâncias mais próximas com um percentual de 10,0% apresentou o melhor resultado (19 ganhos), seguida de

Base	Acurácia										F_1									
	Percentuais selecionados										Percentuais selecionados									
	0,0	0,5	1,0	2,5	5,0	10,0	20,0	40,0	80,0	100,0	0,0	0,5	1,0	2,5	5,0	10,0	20,0	40,0	80,0	100,0
iro	**1,03**	**1,13**	**1,20**	**1,20**	**1,28**	**1,30**	**1,23**	**1,25**	**1,25**	**1,25**	**1,00**	**1,12**	**1,21**	**1,21**	**1,27**	**1,30**	**1,23**	**1,24**	**1,21**	**1,21**
sar	0,96	**1,02**	**1,10**	**1,10**	**1,10**	**1,12**	**1,20**	**1,14**	**1,20**	**1,24**	0,97	**1,04**	**1,12**	**1,13**	**1,14**	**1,15**	**1,25**	**1,18**	**1,25**	**1,29**
ais	0,99	0,98	0,99	0,98	0,98	0,99	0,99	0,99	**1,00**	**1,00**	0,99	0,99	0,99	0,98	0,99	0,99	0,99	0,99	**1,00**	**1,00**
S15	**1,00**	0,99	0,98	0,95	0,96	0,94	0,91	0,88	0,84	0,82	0,99	0,99	0,98	0,96	0,97	0,95	0,92	0,90	0,87	0,85
sem	**1,01**	**1,00**	**1,03**	**1,01**	**1,02**	**1,03**	**1,01**	**1,03**	**1,02**	**1,01**	**1,01**	**1,00**	**1,03**	**1,01**	**1,02**	**1,03**	**1,01**	**1,03**	**1,02**	**1,01**
per	**1,04**	**1,01**	**1,02**	**1,03**	**1,03**	**1,05**	**1,06**	**1,05**	**1,06**	**1,06**	**1,03**	**1,00**	**1,01**	**1,02**	**1,02**	**1,04**	**1,05**	**1,04**	**1,04**	**1,05**
hob	0,98	0,99	**1,00**	0,98	0,96	0,98	0,95	0,96	0,93	0,92	0,98	0,99	**1,00**	0,98	0,96	0,97	0,95	0,96	0,93	0,90
iph	**1,00**	**1,00**	**1,00**	**1,01**	**1,01**	**1,01**	**1,01**	**1,02**	**1,01**	**1,01**	0,99	**1,00**	**1,00**	**1,01**	**1,01**	**1,01**	**1,01**	**1,02**	**1,01**	**1,01**
mov	**1,03**	**1,02**	**1,01**	**1,03**	**1,04**	**1,04**	**1,05**	**1,05**	**1,06**	**1,06**	**1,02**	**1,01**	**1,00**	**1,02**	**1,03**	**1,03**	**1,04**	**1,04**	**1,04**	**1,04**
san	**1,01**	**1,01**	**1,01**	**1,00**	**1,01**	**1,01**	**1,00**	**1,01**	**1,01**	**1,00**	**1,01**	**1,01**	**1,01**	**1,01**	**1,01**	**1,01**	**1,00**	**1,01**	**1,01**	**1,00**
nar	**1,00**	**1,00**	**1,01**	**1,01**	**1,02**	**1,02**	**1,03**	**1,04**	**1,04**	**1,04**	**1,00**	**1,00**	**1,01**	**1,01**	**1,02**	**1,02**	**1,03**	**1,04**	**1,04**	**1,04**
arc	**1,00**	0,99	**1,00**	**1,00**	0,99	0,99	**1,00**	0,99	0,99	0,99	**1,00**	0,99	**1,00**	**1,00**	0,99	0,99	**1,00**	0,99	0,99	0,99
S18	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,01**	**1,01**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,01**	**1,01**	**1,00**	**1,00**
OMD	0,98	0,99	0,99	**1,00**	0,99	**1,00**	0,98	0,98	0,97	0,96	0,98	0,99	0,99	**1,00**	0,99	**1,00**	0,98	0,97	0,96	0,96
HCR	**1,03**	**1,04**	**1,04**	**1,03**	**1,03**	**1,04**	**1,05**	**1,05**	**1,05**	**1,05**	**1,01**	**1,01**	**1,01**	**1,00**	**1,00**	**1,00**	**1,01**	**1,01**	**1,00**	**1,00**
STS	**1,01**	**1,01**	**1,00**	**1,01**	**1,02**	**1,02**	**1,01**	**1,01**	0,99	0,98	**1,00**	**1,01**	**1,00**	**1,01**	**1,02**	**1,02**	**1,01**	**1,01**	**1,00**	0,99
SSt	**1,00**	**1,00**	**1,01**	**1,02**	**1,02**	**1,02**	**1,02**	**1,02**	**1,02**	**1,03**	**1,00**	**1,00**	**1,01**	**1,02**	**1,02**	**1,02**	**1,02**	**1,02**	**1,02**	**1,02**
Tar	**1,00**	**1,00**	**1,00**	**1,01**	**1,00**	**1,01**	**1,00**	0,99	0,99	0,98	**1,00**	**1,00**	**1,00**	**1,01**	**1,00**	**1,01**	**1,00**	0,99	0,99	0,98
Vad	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,01**	**1,00**	**1,00**	**1,01**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,01**	**1,00**	**1,00**	**1,01**	**1,00**
S13	**1,02**	**1,02**	**1,02**	**1,03**	**1,03**	**1,03**	**1,03**	**1,04**	**1,04**	**1,05**	**1,02**	**1,02**	**1,02**	**1,02**	**1,02**	**1,02**	**1,03**	**1,03**	**1,03**	**1,03**
S17	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,01**	**1,01**	**1,01**	**1,01**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,01**	**1,01**	**1,01**	**1,01**	**1,00**
S16	**1,01**	**1,01**	**1,01**	**1,01**	**1,01**	**1,01**	**1,01**	**1,02**	**1,02**	**1,02**	**1,01**	**1,01**	**1,01**	**1,01**	**1,01**	**1,01**	**1,01**	**1,01**	**1,01**	**1,01**
									#Ganhos		16	17	**19**	**19**	17	18	18	16	17	16
									#Melhores		1	0	2	2	0	2	1	**5**	4	**5**
									Ranking		7,05	7,09	6,18	6,00	5,36	4,00	4,50	**3,86**	4,77	5,73

Tabela VI. Ganhos de acurácia e F_1 obtidos com seleção percentual das instâncias da base-fonte mais próximas e mais distantes à base-alvo.

perto por selecionar aleatoriamente com um percentual de 2,5% ou 5,0% (18 ganhos). Levando em consideração o critério da quantidade de vezes que o percentual produziu os melhores resultados, selecionar as mais próximas com um percentual de 20,0% ou as mais próximas e as mais distantes com um percentual de 40,0% obteve os melhores resultados (5 ganhos), sendo que este último percentual também obteve o melhor desempenho no ranking médio (2,82).

Base	Aleatória			Próximas		Próximas e distantes
	2,5%	5,0%	100,0%	10,0%	20,0%	40,0%
iro	**1,13**	**1,18**	**1,19**	**1,18**	**1,22**	**1,24**
sar	**1,11**	**1,10**	**1,27**	**1,16**	**1,23**	**1,18**
ais	0,96	0,97	**1,00**	**1,00**	0,99	0,99
S15	0,96	0,97	0,87	0,93	0,92	0,90
sem	**1,01**	**1,02**	**1,01**	**1,01**	**1,02**	**1,03**
per	**1,02**	**1,04**	**1,05**	**1,06**	**1,04**	**1,04**
hob	0,97	0,97	0,93	0,95	0,96	0,96
iph	**1,03**	**1,00**	0,99	**1,02**	**1,04**	**1,02**
mov	**1,02**	**1,01**	**1,04**	**1,04**	**1,04**	**1,04**
san	**1,01**	**1,00**	**1,00**	**1,00**	**1,01**	**1,01**
nar	**1,02**	**1,02**	**1,04**	**1,02**	**1,03**	**1,04**
arc	**1,00**	**1,00**	0,98	**1,00**	0,99	0,99
S18	**1,00**	**1,01**	**1,00**	**1,00**	**1,01**	**1,01**
OMD	0,99	0,99	0,96	0,98	0,98	0,97
HCR	**1,02**	**1,01**	0,99	**1,00**	**1,00**	**1,01**
STS	**1,01**	**1,01**	**1,00**	**1,00**	**1,01**	**1,01**
SSt	**1,01**	**1,01**	**1,02**	**1,02**	**1,02**	**1,02**
Tar	**1,00**	**1,00**	0,98	**1,00**	**1,00**	0,99
Vad	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**
S13	**1,02**	**1,03**	**1,03**	**1,02**	**1,02**	**1,03**
S17	**1,00**	**1,01**	**1,00**	**1,00**	**1,01**	**1,01**
S16	**1,01**	**1,01**	**1,02**	**1,01**	**1,01**	**1,01**
#Ganhos	18	18	15	**19**	17	16
#Melhores	4	4	4	2	**5**	**5**
Ranking	3,77	3,36	4,05	3,95	2,91	**2,82**

Tabela VII. Comparação entre os melhores percentuais para ganho de F_1 em todas as estratégias.

5. CONCLUSÕES E TRABALHOS FUTUROS

Neste artigo, investigou-se se utilizar dados de bases-fonte proporciona um aumento no desempenho de modelos de classificação para uma base-alvo rotulada, no contexto de análise de sentimentos em tweets. Para isto, foram desenvolvidos dois experimentos, o primeiro que agrega a totalidade da união das bases-fonte ao conjunto de treinamento do classificador e o segundo propondo estratégias de seleção de instâncias dessa união de bases-fonte de acordo com três estratégias: (I) seleção aleatória de

instâncias, (II) seleção das instâncias mais próximas a cada instância das partições de treinamento da base-alvo e (III) seleção das instâncias mais próximas e mais distantes de cada instância das partições de treinamento da base-alvo. Para todos os experimentos, o conjunto de treinamento era balanceado e os modelos gerados foram testados em partições da base-alvo por meio de validação cruzada. Os resultados foram comparados com o desempenho do classificador treinado apenas com a base-alvo em termos de acurácia e F_1 ponderado por intermédio do cálculo do ganho – divisão entre os valores das métricas usando a união de bases-fonte com a base-alvo e usando somente a base-alvo.

Os resultados do primeiro experimento mostraram que aproveitar um conjunto de bases-fonte para compor o conjunto de treinamento produz ganhos de desempenho para a maioria das bases-alvo, tanto em termos de acurácia quanto em termos de F_1. No entanto, esse ganho não se mostrou elevado para a maioria das bases, o que indicou que alguma estratégia de seleção de instâncias poderia ser útil.

Os resultados do segundo experimento apontaram que algumas combinações de estratégia e percentual apresentaram bom desempenho. Para a seleção aleatória, os melhores desempenhos ocorreram selecionando 2,5%, 5,0% e 100,0%. Considerando a seleção das instâncias mais próximas a cada instância das partições de treinamento da base-alvo, selecionar 10,0% ou 20,0% obteve o melhor resultado, ao passo que para a estratégia que inclui selecionar também as mais distantes os melhores resultados foram encontrados com a seleção de 40,0% da união das bases-fonte. Entre essas seis combinações de estratégia-percentual, a que apresentou o melhor resultado geral foi a seleção de 40,0% com a estratégia das mais próximas e das mais distantes. Estes resultados indicam que utilizar uma união de bases-fonte para ser agregada ao conjunto de treinamento de classificadores para uma base-alvo pode trazer ganhos de desempenho, em particular porque esse conjunto de bases-fonte pode ser usado para ampliar, balancear e diversificar o conjunto de treinamento.

Trabalhos futuros incluem a utilização de outras métricas para a seleção de instâncias. Adicionalmente, um parâmetro que regule a proporção de instâncias mais próximas e mais distantes a serem utilizadas pode ser acrescentado e ajustado.

REFERENCES

Barreto, S., Moura, R., Carvalho, J., Paes, A., and Plastino, A. Sentiment analysis in tweets: an assessment study from classical to modern text representation models. *CoRR* vol. abs/2105.14373, 2021.

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

Carvalho, J. and Plastino, A. On the evaluation and combination of state-of-the-art features in twitter sentiment analysis. *Artificial Intelligence Review* vol. 54, pp. 1887–1936, 03, 2021.

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

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

Liu, M., Song, Y., Zou, H., and Zhang, T. Reinforced training data selection for domain adaptation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. ACL, pp. 1957–1968, 2019.

Martínez-Cámara, E., Martín-Valdivia, M., López, L., and Montejo-Ráez, A. Sentiment analysis in twitter. *Natural Language Engineering* vol. 20, pp. 1–28, 01, 2014.

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

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* vol. 12, pp. 2825–2830, 2011.

Ruder, S., Ghaffari, P., and Breslin, J. G. Data selection strategies for multi-domain sentiment analysis. *CoRR* vol. abs/1702.02426, 2017.

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