UNIVERSIDADE FEDERAL FLUMINENSE

ELISEU PAZ E SILVA DE GUIMARÃES

SELEÇÃO DE DADOS PARA TRANSFERÊNCIA DE APRENDIZADO NO CONTEXTO DE ANÁLISE DE SENTIMENTOS EM TWEETS

NITERÓI

2021

UNIVERSIDADE FEDERAL FLUMINENSE

ELISEU PAZ E SILVA DE GUIMARÃES

SELEÇÃO DE DADOS PARA TRANSFERÊNCIA DE APRENDIZADO NO CONTEXTO DE ANÁLISE DE SENTIMENTOS EM TWEETS

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Computação da Universidade Federal Fluminense como requisito parcial para a obtenção do Grau de Mestre em Computação. Área de concentração: Ciência da Computação

Orientador:
ALEXANDRE PLASTINO DE CARVALHO

Coorientadora:
ALINE MARINS PAES CARVALHO

NITERÓI

2021

Ficha catalográfica automática - SDC/BEE
Gerada com informações fornecidas pelo autor

G963s Guimarães, Eliseu Paz e Silva de
SELEÇÃO DE DADOS PARA TRANSFERÊNCIA DE APRENDIZADO NO CONTEXTO DE ANÁLISE DE SENTIMENTOS EM TWEETS / Eliseu Paz e Silva de Guimarães ; ALEXANDRE PLASTINO DE CARVALHO, orientador ; ALINE MARINS PAES CARVALHO, coorientadora. Niterói, 2021.
75 f.

Dissertação (mestrado)-Universidade Federal Fluminense, Niterói, 2021.

DOI: http://dx.doi.org/10.22409/PGC.2021.m.11447293789

1. Aprendizado de Máquina. 2. Produção intelectual. I. CARVALHO, ALEXANDRE PLASTINO DE, orientador. II. CARVALHO, ALINE MARINS PAES, coorientadora. III. Universidade Federal Fluminense. Instituto de Computação. IV. Título.

CDD -

Bibliotecário responsável: Debora do Nascimento - CRB7/6368

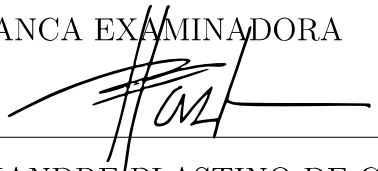
ELISEU PAZ E SILVA DE GUIMARÃES

SELEÇÃO DE DADOS PARA TRANSFERÊNCIA DE APRENDIZADO NO CONTEXTO DE ANÁLISE DE SENTIMENTOS EM TWEETS

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Computação da Universidade Federal Fluminense como requisito parcial para a obtenção do Grau de Mestre em Computação. Área de concentração: Ciência da Computação

Aprovada em Agosto de 2021.

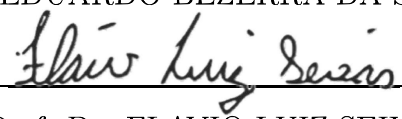
BANCA EXAMINADORA



Prof. Dr. ALEXANDRE PLASTINO DE CARVALHO
Orientador, UFF

Profa. Dra. ALINE MARINS PAES CARVALHO
Coorientadora, UFF

Prof. Dr. EDUARDO BEZERRA DA SILVA, CEFET/RJ



Prof. Dr. FLAVIO LUIZ SEIXAS, UFF

Niterói
2021

"We can only see a short distance ahead, but we can see plenty there that needs to be done."

Alan Turing

Dedico esta dissertação ao meu filho Antônio, à Dilza, minha esposa, e à Ana Lúcia, minha mãe. Obrigado ontem, hoje e sempre.

Agradecimentos

Agradeço, com respeito e admiração, aos meus orientadores, Alexandre Plastino e Aline Paes, pela condução desta pesquisa, feita com ciência e compreensão. Nada que aqui está seria possível sem vocês. Ao grupo de pesquisa, agradeço aos conselhos sempre precisos, em especial ao Jonn e à Dani, fundamentais para os artigos escritos. Agradeço, ainda, aos professores e funcionários do IC, em particular à profa. Flávia Bernardini e ao prof. Celso, cujas disciplinas e conselhos tanto contribuíram para a minha formação.

Agradeço ao meu filho Antônio, alegria maior dos meus dias nesses dois anos de mestrado e à minha esposa Dilza pelos amor, carinho, apoio, companheirismo e puxões de orelha científicos. À minha mãe, Ana Lúcia, por ser, há muito tempo e para sempre, o meu Norte. Às minhas irmãs, Anne e Lu, pela amizade e carinho que crescem com o tempo e pelo apoio a todo momento. Ao meu tio Paulo e às minhas tias Leca, Sílvia e Cláudia, pelo carinho e apoio constantes.

Agradeço aos meus amigos Fernanda e Rodrigo, pela amizade de tantos anos e pela certeza de que, mesmo em tempos tão difíceis, é possível se fazer presente de alguma forma. À Roberta Valentim, para quem o mestrado significou o fim do contato diário, agradeço pelo carinho e presença. Obrigado, minha "terceira irmã". À Marcela Trindade, que acompanhou o mestrado desde antes de ele existir, agradeço pelo carinho e pelo incentivo inestimáveis. Agradeço ao Wilen, amigo e médico, pela amizade e pela condução precisa e carinhosa do meu quadro quando tive COVID.

Agradeço à UFF, por proporcionar um curso de tão alta qualidade, mesmo em uma situação tão adversa do mundo. À UFRJ, por ter sido o início de minha formação acadêmica, nos tempos de graduação. E à Marinha, pela liberação para realizar este mestrado.

Por fim, e definitivamente não menos importante, agradeço ao povo brasileiro. Todos aqueles que, com o suor de seu trabalho, possibilitaram esta pesquisa, pagando meus salários e financiando a UFF. Agradeço, ainda, a todos que lutam pela universidade pública. Para que ela continue a ser uma força transformadora da sociedade.

Resumo

Com o advento e a popularização das redes sociais, cada vez mais pessoas sentem-se livres para expressarem suas opiniões sobre assuntos variados naqueles ambientes. Esse tipo de atitude gera um volume crescente de dados, cuja análise constitui importante ferramenta no processo de tomada de decisão de instituições, governos ou pessoas, que podem aferir seu desempenho em relação a um público-alvo desejado. O campo de estudo computacional que visa a atender este objetivo é a análise de sentimentos, que tem a classificação de polaridade de textos como uma de suas tarefas de maior destaque. Para atender à necessidade de classificar textos como positivos ou negativos, destaca-se o uso de abordagens baseadas em aprendizado de máquina supervisionado, nas quais um classificador é treinado com um conjunto de dados de um determinado domínio cujos rótulos (positivos ou negativos) são conhecidos. A ideia por trás dessa abordagem é que este classificador seja capaz de predizer os rótulos de novos dados deste mesmo domínio. No entanto, dados rotulados nem sempre estão disponíveis, pois o domínio de interesse pode ser raro e ter dados escassos, ou ainda rotular manualmente os dados pode ser proibitivo. Nesse cenário, surgem estratégias de transferência de aprendizado, que buscam aproveitar o conhecimento adquirido em um determinado domínio-fonte para adaptar ou reusar classificadores para um determinado domínio-alvo. Uma das abordagens utilizadas se baseia na seleção ou enriquecimento de dados a partir de um domínio-fonte, o que tem sido amplamente proposto na literatura. No entanto, há carência de estudos específicos para a seleção de instâncias no desafiador cenário do Twitter. Esta dissertação se propõe a investigar técnicas de seleção de dados para transferência de aprendizado no contexto de análise de sentimentos em tweets. Para isso, são realizados experimentos utilizando um conjunto de 22 bases de dados de tweets em inglês. Nestes experimentos, são propostas técnicas: (i.) de seleção de bases-fonte para treinar classificadores para uma base-alvo não-rotulada, (ii.) de seleção de instâncias da união das bases-fonte para treinar classificadores para uma base-alvo não-rotulada e (iii.) de seleção de instâncias da união das bases-fonte para treinar classificadores para uma base-alvo rotulada. Com as técnicas propostas, observa-se que o tamanho do conjunto de treinamento desempenha um papel fundamental na capacidade preditiva dos classificadores e que utilizar conjuntos de treinamento balanceados e diversos constitui-se uma boa decisão para os métodos de transferência de aprendizado que se baseiam em seleção de instâncias e reuso de classificadores.

Palavras-chave: Análise de sentimentos; Transferência de aprendizado; Seleção de dados; Aprendizado de Máquina; Twitter

Abstract

The advent and popularization of social networks have been leading more and more people to feel free to express their opinions on various issues in those environments. This type of attitude generates a growing volume of data, whose analysis is an important tool in the decision-making process of institutions, governments or people, that can assess their performance related to a desired target audience. The computational field of study that aims to meet this objective is called sentiment analysis, which has the polarity classification of texts as one of its most prominent tasks. To meet the need to classify texts as positive or negative, the use of approaches based on supervised machine learning is promising, in which a classifier is trained with a dataset from a given domain whose labels (positive or negative) are known. The idea behind this approach is that this classifier can predict the labels of new data from this same domain. However, labeled data are not always available as the domain of interest can be rare and data scarse, or manually labeling the data can be prohibitive. In this scenario, transfer learning strategies arise, seeking to take advantage of the knowledge acquired in a given source domain to adapt or reuse classifiers for a given target domain. One of the approaches used is based on data selection or enrichment from a source domain - which has been widely proposed in the literature. However, there is a lack of specific studies for instance selection in the challenging scenario of Twitter. This dissertation seeks to investigate data selection techniques for transfer learning in the scenario of sentiment analysis in tweets. For this, experiments are conducted using a set of 22 tweets datasets in English. These experiments propose techniques: (i.) to select source datasets to train classifiers for an unlabeled target dataset, (ii.) to select instances of the union of source datasets to train classifiers for an unlabeled target dataset and (iii.) to select instances of the union of source datasets to train classifiers for a labeled target dataset. With the proposed techniques, it is observed that the size of training set plays a fundamental role in the predictive capability of the classifiers and that using balanced and diverse training sets constitutes a good decision for transfer learning methods based on instance selection and reuse of classifiers.

Keywords: Sentiment analysis; Transfer learning; Data selection; Machine Learning; Twitter

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

BoW	:	*Bag-of-Words*.
CBoW	:	*Continuous Bag-of-Words*.
PLN	:	Processamento de Linguagem Natural.
RWMD	:	*Relaxed Word Moving Distance*.
SVM	:	*Support Vector Machines*.
TF-IDF	:	*Term Frequency–Inverse Document Frequency*.
w2v	:	*word2vec*.

Sumário

Capítulo 1

Introdução

A análise de sentimentos pode ser definida como o estudo computacional de opiniões, sentimentos, emoções, humores e atitudes das pessoas [19]. Dentro do campo da análise de sentimentos, destaca-se a classificação de polaridade de textos, que se caracteriza por atribuir a textos uma posição favorável (classe positiva) ou desfavorável (classe negativa).

A classificação de polaridade permite que instituições dos mais variados tipos (empresas, organizações não-governamentais, órgãos governamentais, etc) possam aferir o seu desempenho frente a um público-alvo, servindo de importante ferramenta no processo de tomada de decisão, além de permitir uma análise rápida do impacto de suas ações neste público. Com o avanço e a popularização das redes sociais nos últimos anos, há uma grande quantidade de pessoas que se sentem livres para opinar sobre os mais diversos assuntos, o que gera um volume crescente de dados a serem analisados. Dentre estas redes sociais podemos destacar o Twitter[1].

O Twitter é um serviço de microblog de textos curtos, chamados tweets, limitados atualmente a 280 caracteres, mas cujo limite era de 140 caracteres até novembro de 2017. A análise de sentimentos em tweets é uma tarefa desafiadora, visto que a informalidade presente nesses textos leva a uso incorreto de gramática, presença de palavras escritas de forma errada e falta de contexto [22].

Há duas abordagens principais que são adotadas para enfrentar o problema de classificação de polaridade: métodos baseados em léxicos e estratégias de aprendizado de máquina. A primeira abordagem busca inferir a polaridade de textos a partir do uso de léxicos previamente anotados com o sentimento das palavras ou frases que os compõem. A segunda abordagem, por sua vez, extrai características de textos rotulados de

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

um determinado domínio, ou seja, textos cujas classes já são conhecidas, e utiliza essas características como atributos para o treinamento de classificadores que busquem predizer a polaridade de outros textos desse mesmo domínio.

Mais formalmente, considere um determinado domínio $\mathfrak{D}$ e um conjunto de dados $\mathbf{D} = \{(\mathbf{x_i}, y_i) \mid \mathbf{x_i} = (x_{i1}, x_{i2}, ..., x_{ik}), 1 \leq i \leq n\}$, formado por n instâncias pertencentes a este domínio, cada uma com k atributos e uma classe y_i associada, $y_i \in \{0, 1\}$, com 0 representando a classe negativa e 1 a classe positiva. Uma abordagem clássica de aprendizado de máquina para a classificação de polaridade usará estas instâncias para treinar uma função de predição $h(\mathbf{x})$, de forma a minimizar uma função de perda $l(h(\mathbf{x}), y)$. A ideia é que esta função $h(\mathbf{x})$ seja capaz de predizer a classe y de novas instâncias deste domínio.

Contudo, nem sempre há dados rotulados suficientes em um domínio de interesse para que sejam gerados classificadores adequados, ou seja, classificadores cuja função de predição minimize a função de perda satisfatoriamente. Isso pode decorrer tanto do fato do domínio ser raro, e a quantidade de dados disponíveis ser escassa, quanto do fato de ser custoso rotular manualmente os dados existentes. Neste contexto, uma possibilidade é a adoção de técnicas de transferência de aprendizado, que aproveita conhecimento obtido em um determinado domínio, chamado de domínio-fonte, para adaptá-lo ou reusá-lo em outro domínio, chamado de domínio-alvo [27]. Dentro deste tipo de abordagem, uma possibilidade é utilizar dados rotulados do domínio-fonte para treinar um classificador para o domínio-alvo [40]. Em geral, esses domínios são relacionados, porém com distribuições distintas. Esta abordagem permite enfrentar tanto o problema de escassez de dados do domínio-alvo, utilizando o domínio-fonte para enriquecer o conjunto de treinamento, quanto o problema da inviabilidade de rotular os dados do domínio-alvo, usando o domínio-fonte como conjunto de treinamento. Recentemente, o uso de transferência de aprendizado em PLN tem sido um campo de estudo bastante atrativo [33].

Em comparação com a abordagem tradicional de aprendizado de máquina realizada dentro de um mesmo domínio, considere os domínios-fonte e alvo $\mathfrak{D}_\mathfrak{s}$ e $\mathfrak{D}_\mathfrak{t}$, respectivamente. Para a situação em que há escassez de dados rotulados do domínio-alvo, o conjunto de dados de n instâncias $\mathbf{D_t} = \{(\mathbf{x_{ti}}, y_{ti}), \mid \mathbf{x_{ti}} = (x_{ti1}, x_{ti2}, ..., x_{tik}), 1 \leq i \leq n\}$, pertencente a este domínio, pode não possuir elementos suficientes para treinar um estimador $h_t(\mathbf{x_t})$ que minimize de forma satisfatória a função de perda $l(h_t(\mathbf{x_t}), y_t)$. Desta forma, pode se fazer uso de um conjunto de dados, pertencente ao domínio-fonte, $\mathbf{D_s} = \{(\mathbf{x_{sj}}, y_{sj}), \mid \mathbf{x_{sj}} = (x_{sj1}, x_{sj2}, ..., x_{sjk}), 1 \leq j \leq m\}$, com m instâncias, de modo que seja treinada uma função

de predição $h_{st}(\mathbf{x_{st}}), \mathbf{x_{st}} \in \mathbf{D_s} \bigcup \mathbf{D_t}$ que minimize a função de perda $l(h_{st}(\mathbf{x_{st}}), y_{st})$. Para o caso em que é inviável rotular dados do domínio-alvo, o conjunto de dados deste domínio não possuirá as classes das instâncias, ou seja, $\mathbf{D_t} = \{(\mathbf{x_{ti}}), \mid \mathbf{x_{ti}} = (x_{ti1}, x_{ti2}, ..., x_{tik}), 1 \leq i \leq n)\}$. Assim, é treinada a função de predição $h_s(\mathbf{x_s})$, que minimiza a função de perda $l(h_s(\mathbf{x_s}), y_s)$. Tanto na situação em que há escassez de dados rotulados no domínio-alvo quanto naquela em que é inviável rotulá-los, espera-se que a função de predição seja capaz de classificar corretamente novos exemplos que pertençam ao domínio-alvo $\mathfrak{D}_t$.

Diante da abundância e variedade de dados rotulados de diversos outros domínios, surge, então, o problema de como selecionar os dados que, de fato, serão usados como parte ou totalidade do conjunto de treinamento. Ou seja, passa a haver a necessidade de se estabelecerem critérios para que o conjunto de treinamento possa ser formado para a geração de um classificador com bom desempenho no contexto em que os dados de um domínio-alvo não são suficientes para gerá-lo. Com o objetivo de solucionar esta questão, diversos trabalhos têm proposto abordagens variadas [13, 15, 17, 18, 20, 30, 31, 34, 35, 38, 42, 43] para diferentes tarefas. Essas abordagens incluem aprender métricas de similaridade, usar critérios de seleção baseados em diversidade, fazer combinações entre métricas de similaridade e atribuir pesos às instâncias ou aos classificadores gerados. As abordagens propostas por este estudo visam investigar o uso de métricas de similaridade, incluindo a dissimilaridade como critério de escolha, sem a necessidade de aprendizado de métricas ou de atribuições de pesos aos dados. Além disso, este trabalho utiliza um extenso conjunto de bases de dados de tweets, colocando seu foco nesse cenário desafiador para a análise de sentimentos.

Nesse contexto, esta dissertação pretende contribuir com abordagens para a seleção de dados para transferência de aprendizado no contexto de análise de sentimentos em tweets. Uma vez que essas abordagens e suas avaliações foram reportadas em três artigos distintos – um publicado e dois submetidos – resolvemos trazer esses artigos em três apêndices e resumir os seus respectivos resultados e contribuições. Os artigos são os seguintes:

1. Guimarães, Eliseu; Carvalho, Jonnathan; Paes, Aline; Plastino, Alexandre. Transfer learning for Twitter sentiment analysis: Choosing an effective source dataset. VIII Symposium on Knowledge Discovery, Mining and Learning (KDMiLe), 161-168, 2020.

2. Guimarães, Eliseu; Carvalho, Jonnathan; Paes, Aline; Plastino, Alexandre. Exploring model transfer strategies for sentiment analysis in Twitter. Submetido ao XVIII Encontro Nacional de Inteligência Artificial e Computacional (ENIAC), 2021.

3. Guimarães, Eliseu; Vianna, Daniela; Paes, Aline; Plastino, Alexandre. Enriching datasets for sentiment analysis in tweets with instance selection. Submetido ao IX Symposium on Knowledge Discovery, Mining and Learning (KDMiLe), 2021.

O primeiro artigo analisa o desempenho de diversas métricas de similaridade aplicadas entre bases de dados para a seleção de uma base-fonte com o intuito de treinar um classificador a ser aplicado a uma base-alvo. Neste artigo, além das métricas, são avaliadas variações nas formas de representação dos dados. Adicionalmente, é realizado um experimento com o objetivo de verificar, entre as bases-fonte disponíveis, quais bases geram os melhores classificadores para cada base-alvo, independentemente da métrica de similaridade escolhida. Os resultados apontam que a utilização de similaridade de cosseno com a representação usando *Bag-of-Words* (BoW) [37] com *term frequency–inverse document frequency* (TF-IDF) obteve o melhor resultado. Além disso, os classificadores treinados em algumas das bases utilizadas obtiveram um bom desempenho preditivo para diversas bases-alvo, o que indica que essas bases podem ser boas bases-fonte para o treinamento de modelos de classificação. Este artigo foi publicado no simpósio KDMiLe 2020.

O segundo artigo explora estratégias de transferência de modelos para a análise de sentimentos em tweets. Usando um conjunto de bases-fonte, são treinados classificadores para uma base-alvo não-rotulada, inicialmente unindo as bases-fonte disponíveis e, posteriormente, propondo estratégias de seleção de dados desta união. Este estudo mostrou que a união das bases-fonte como conjunto de treinamento produz um bom desempenho, que pode ser levemente melhorado com uma seleção de instâncias que considere similaridade e dissimilaridade, além de balanceamento. Este artigo foi submetido ao ENIAC 2021, estando atualmente em processo de revisão.

O terceiro artigo aborda o caso em que a base-alvo é rotulada e as bases-fonte são utilizadas com o objetivo de enriquecer o conjunto de treinamento. Nesse estudo, também são testados tanto o uso da união de um conjunto de bases-fonte, quanto a seleção de instâncias dessa união, sendo sempre realizado o balanceamento dos conjuntos de treinamento. Os resultados mostram que utilizar critérios de seleção baseados em similaridade e dissimilaridade produzem aumento no desempenho dos classificadores gerados, já a partir da seleção de um percentual pequeno da união das bases-fonte. Este artigo foi submetido ao KDMiLe 2021 e está no momento em processo de revisão.

O restante desta dissertação está organizado da seguinte maneira. No Capítulo 2 são apresentadas as contribuições das pesquisas realizadas, resumindo os artigos dos apêndices. Mais precisamente, são apresentadas breves descrições de alguns trabalhos rela-

cionados às pesquisas desenvolvidas e é descrita a metodologia geral utilizada nos três artigos. São também mostradas de forma resumida as metodologias específicas de cada artigo, além de serem elencados seus principais resultados e conclusões. Por sua vez, no Capítulo 3 são apresentadas as conclusões gerais desta dissertação, assim como as possibilidades vislumbradas de trabalhos futuros.

Capítulo 2

Contribuições

O objetivo deste capítulo é apresentar as contribuições desta dissertação. Para isso, são resumidos os três artigos resultantes da exploração de técnicas de seleção de dados para transferência de aprendizado no contexto de análise de sentimentos em tweets. Especificamente, na Seção 2.1 são apresentadas breves descrições de alguns trabalhos relacionados a estes artigos, enquanto na Seção 2.2, a metodologia geral utilizada nos três artigos é descrita. A Seção 2.3 traz o resumo do artigo do Apêndice A, apresentando de forma resumida sua metodologia específica, seus resultados e as principais conclusões. Na Seção 2.4 encontra-se o resumo do artigo do Apêndice B, com sua metodologia, seus resultados e suas conclusões, o mesmo ocorrendo para o artigo do Apêndice C, cujo resumo é apresentado na Seção 2.5.

2.1 Trabalhos relacionados

Nesta seção, serão descritos alguns trabalhos que tratam o problema de seleção de dados para transferência de aprendizado e as abordagens propostas para solucioná-lo.

Em [18], o problema da dificuldade de se rotular os dados do domínio-alvo é tratado propondo-se um algoritmo baseado em regressão logística que, para treinar o classificador utilizando uma base-fonte e uma base-alvo parcialmente rotulada, usa um termo de compatibilidade entre cada uma das instâncias da base-fonte e a base-alvo. Este termo controla a participação de cada instância da base-fonte no treinamento do classificador e reflete a incompatibilidade entre essas instâncias e a base-alvo. Os resultados utilizando uma base-alvo artificial e uma base-alvo sobre câncer de mama mostram que o algoritmo proposto apresenta melhor desempenho do que a abordagem tradicional da regressão logística e que a utilização do termo de compatibilidade para atribuição de peso às instâncias

possui relevância para o problema estudado. Nos artigos oriundos desta dissertação, e que estão nos Apêndices A e B, é usado o algoritmo de regressão logística, sem atribuição de pesos às instâncias, o que simplifica o processo de treinamento. Cabe ressaltar que a utilização de pesos poderia aumentar a acurácia do classificador.

Em [15], três parâmetros associados a instâncias são calculados com o objetivo de encontrar a distribuição de probabilidade de uma base-alvo parcialmente rotulada. Estes parâmetros levam em consideração: as probabilidades de uma instância da base-fonte ser classificada corretamente usando as distribuições de probabilidade da base-alvo e da base-fonte; as probabilidades de ocorrência dos atributos das instâncias da base-fonte na base-fonte e na base-alvo; e a probabilidade de as instâncias não-rotuladas da base-alvo pertencerem a cada uma das classes do problema. Outros três parâmetros globais são utilizados para controlar a contribuição de cada um dos métodos de aproximação utilizados a fim de calcular o argumento que maximiza uma função de verossimilhança. O trabalho mostra que considerar a divergência entre os domínios-alvo e fonte aumenta o desempenho da classificação, além de mostrar que utilizar informação da base-alvo é mais importante do que eliminar instâncias enganosas da base-fonte. Assim como nesse trabalho, no artigo do Apêndice C, as instâncias da base-alvo são utilizadas para o treinamento, assumindo a base-alvo totalmente rotulada, mas não são utilizados parâmetros para atribuição de pesos, nem são calculadas medidas de divergência entre as bases. Para a seleção de instâncias, nos três artigos desenvolvidos, são utilizadas métricas de similaridade ou distância entre bases, entre bases e instâncias, ou entre instâncias.

O trabalho desenvolvido por [31] propõe uma abordagem baseada em duas métricas: similaridade de domínio e complexidade de domínio. A similaridade de domínio calcula a divergência entre os domínios como forma de aferir sua semelhança, ao passo que a complexidade de domínio é calculada como uma medida de autodivergência de um domínio. Essas métricas são usadas no cálculo de um fator que determina o quanto da base-fonte será usada no treinamento em conjunto com a base-alvo rotulada. Os artigos dos Apêndices B e C oriundos desta dissertação, embora não utilizem fatores calculados com base em divergências dos domínios, também consideram a possibilidade de utilizar apenas parte das bases-fonte para o treinamento, sendo que apenas no artigo do Apêndice C a base-alvo é assumida como rotulada.

Em [43], são selecionadas instâncias de várias bases-fonte para compor o conjunto de treinamento para uma base-alvo não-rotulada, assim como ocorre no artigo do Apêndice B. Nesse trabalho, são utilizados *reviews* da Amazon, enquanto nesta dissertação

são utilizados tweets, o que coloca ambos os trabalhos no campo do PLN. Em [43], faz-se a seleção por intermédio de *PU learning* (Positive Unlabeled learning), enquanto o artigo do Apêndice B utiliza métricas de distância. São propostas duas abordagens para formar o conjunto de treinamento: seleção de instâncias e atribuição de pesos. Na primeira, instâncias da base-fonte que possuem maior probabilidade de pertencerem à base-alvo são selecionadas como dados de treinamento. Na segunda, são utilizados pesos no treinamento, oriundos de uma calibração na probabilidade de as instâncias pertencerem à base-alvo. Em [17], utilizam-se os *reviews* da Amazon como bases-fonte rotuladas e base-alvo não rotulada. A ideia principal desse trabalho é que as instâncias da base-alvo que estejam mais próximas da base-fonte têm maior probabilidade de serem classificadas corretamente, com um classificador treinado na base-fonte. São, então, atribuídos pesos, representando esses graus de confiança, às instâncias da base-alvo e é feita uma regularização, de forma que os rótulos se propaguem suavemente ao longo da base-alvo. Embora os artigos dos Apêndices B e C também sigam a ideia de selecionar as instâncias mais próximas, eles também consideram a possibilidade de selecionar as instâncias mais distantes e fazem treinamentos sem atribuições de pesos.

Em [34], é realizada uma análise de estratégias de seleção de dados, levando em consideração três fatores importantes: a representação dos dados, a métrica de similaridade e o nível de seleção. Foram utilizadas três representações nas avaliações, cada uma delas associada à métrica de similaridade que mais frequentemente é utilizada com a representação. No que diz respeito ao nível de seleção, são consideradas as hipóteses de se medir a similaridade da base-alvo com as bases-fonte inteiras, com as instâncias das bases-fonte individualmente e com subconjuntos das bases-fonte. Os resultados obtidos indicaram que utilizar a seleção no nível dos subconjuntos pode ser uma melhor opção do que selecionar no nível de instâncias. No artigo do Apêndice A, também são consideradas variações nas representações das bases para o cálculo das métricas de similaridade. São cinco formas de representação sendo que, para duas delas, são utilizadas duas métricas de similaridade e, para as outras três, apenas uma. No entanto, nos artigos apresentados nesta dissertação, não são consideradas as medições de similaridade no nível de subconjuntos, apenas no nível de bases-fonte inteiras (artigos dos Apêndices A e B) ou de instâncias das bases-fonte (artigos dos Apêndices B e C).

O trabalho desenvolvido em [13] apresenta uma abordagem do tipo *mixture-of-experts*, que se apoia na ideia que diferentes bases-fonte estão alinhadas a diferentes regiões da base-alvo. Diante disso, é aprendida uma métrica do tipo *point-to-set*, calculada entre as instâncias da base-alvo e as bases-fonte e utilizada para ponderar os resultados de

classificadores treinados com essas bases-fonte. Os resultados do estudo apontam que a técnica proposta obtém melhores resultados em termos de acurácia do que utilizar como conjunto de treinamento apenas uma das bases-fonte ou mesmo a união de todas as bases-fonte sem ponderação. Os artigos desenvolvidos para esta dissertação consideram métricas do tipo *point-to-set* (Apêndices B e C), mas não fazem atribuições de pesos, nem utilizam mais de um tipo de classificador para a predição, o que torna os métodos adotados nesta dissertação mais simples. Novamente, cabe ressaltar que pesos ou a utilização de mais classificadores poderiam ter um impacto positivo na acurácia.

Em [20], é apresentada uma abordagem de aprendizado por reforço usando um *framework* composto de dois módulos que, simultaneamente, busca instâncias relevantes na base-fonte e aprende melhores representações para as instâncias. Um dos módulos seleciona os dados levando em consideração um vetor de distribuição gerado na seleção de dados do passo anterior, enquanto o outro extrai os atributos dos dados, atualiza recompensas com o intuito de gerar o vetor de distribuição e gera um classificador específico para a tarefa. Os resultados mostraram que esta abordagem teve um melhor desempenho em três de quatro bases-alvo utilizadas, em comparação com resultados de outros estudos. Ao contrário desse artigo, nesta dissertação, não são utilizados métodos de realimentação, simplificando o treinamento dos classificadores.

Por sua vez, [42] calcula a correlação entre métricas de similaridade e perda de acurácia na transferência de aprendizado entre bases compostas de textos de livros e periódicos. São calculadas seis métricas de similaridade, computadas usando as frequências relativas das palavras, e a divergência de Rényi apresenta a mais alta correlação. Em [30], foram utilizadas seis tipos de métricas e dois tipos de representação de atributos para a tarefa de *parsing* em conjuntos de dados de texto em inglês e em holandês. O melhor resultado obtido foi utilizando modelagem por tópicos e métrica variacional. Em [35], foi proposta uma abordagem que aprendesse métricas de seleção de dados por intermédio de otimização Bayesiana. Considerando três tipos de representação dos dados, foram usados como atributos para o aprendizado da métrica de seleção seis métricas de similaridade, além de métricas para aferir a diversidade intrínseca à base-fonte. Para a análise de sentimentos foram usados dados de *reviews* da Amazon [4] e os resultados tiveram um melhor desempenho do que as métricas existentes à época. Em [38], foram utilizadas três métricas de distância, tendo como base-alvo um conjunto de tweets no contexto das eleições presidenciais no Brasil em 2018 e como bases-fonte conjuntos de dados provenientes de mídias sociais. Concluiu-se que usar bases-fonte mais similares à base-alvo é uma estratégia melhor e que a utilização de bases-fonte dissimilares reduz o desempenho dos classifica-

dores. Todos estes trabalhos, assim como esta dissertação, se inserem no campo do PLN, além de utilizarem mais de uma métrica de similaridade para atingirem seus objetivos, assim como foi feito no artigo do Apêndice A. Diferentemente do realizado em [42], no desenvolvimento deste trabalho, não foram calculadas correlações entre as métricas e a perda de acurácia, assim como não foram utilizados métodos de aprendizado de métricas, como feito em [35]. Semelhantemente ao feito em [38], as bases-alvo desta dissertação são compostas por tweets, embora nesse trabalho as bases-fonte fossem provenientes também de outras redes sociais e, nesta dissertação, todas têm origem no Twitter.

2.2 Metodologia geral

Esta seção aborda a metodologia em comum utilizada em todos os experimentos conduzidos neste trabalho.

2.2.1 Bases de dados

Nesta dissertação, são utilizadas 22 bases de dados compostas de tweets sobre domínios variados[1] [6]. A Tabela 2.1 apresenta as bases, seus domínios [2], suas quantidades de instâncias positivas e negativas, com seus respectivos percentuais e a quantidade total de instâncias de cada base. Além disso, são apresentadas as abreviações que são usadas para as bases de dados.

2.2.2 Pré-processamento

Como pré-processamento, inicialmente os tweets tiveram suas menções a usuários e URLs substituídas por expressões únicas. Em seguida, todas as letras foram colocadas em minúsculas e os tweets foram tokenizados. A Figura 2.1 apresenta de forma esquemática os passos adotados, assim como o resultado de cada etapa de processamento.

Para a extração dos atributos dos tweets foram utilizados *word embeddings*, que são representações distribuídas para as palavras, aprendidas a partir de um *corpus* linguístico [3]. Neste tipo de representação, uma rede neural é treinada de forma a aprender as relações entre as palavras e posicioná-las em um espaço vetorial cuja dimensionalidade é baixa quando comparada à dimensionalidade de modelos vetoriais anteriores, como por exemplo o BoW. Nos *word embeddings*, cada coordenada do vetor que representa uma

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

Base	Abreviações	Domínio	Positivas	Negativas	Total
irony [11]	iro	Ironia	22 (34%)	43 (66%)	65
sarcasm [11]	sar	Sarcasmo	33 (46%)	38 (54%)	71
aisopos[2]	ais/ntu	Genérico	159 (57%)	119 (43%)	278
SemEval15-Fig[3]	S15	Ironia/metáforas	47 (15%)	274 (85%)	321
sentiment140 [10]	sem/stm	Genérico	182 (51%)	177 (49%)	359
person [7]	per	Pessoas	312 (71%)	127 (29%)	439
hobbit [21]	hob	Filmes	354 (68%)	168 (32%)	522
iphone [21]	iph	Produtos	371 (70%)	161 (30%)	532
movie [7]	mov	Filmes	460 (82%)	101 (18%)	561
sanders[4]	san	Negócios	570 (47%)	654 (53%)	1224
Narr [26]	nar/Nar	Genérico	739 (60%)	488 (40%)	1227
archeage [21]	arc	Jogos	724 (42%)	994 (58%)	1718
SemEval18 [24]	S18	Equity Evaluation Corpus	865 (47%)	994 (53%)	1859
OMD [8]	OMD/deb	Debate presidencial	710 (37%)	1196 (63%)	1906
HCR [39]	HCR	Reforma do sistema de saúde	539 (28%)	1369 (72%)	1908
STS-gold [36]	STS	Genérico	632 (31%)	1402 (69%)	2034
SentiStrength [41]	SSt	Genérico	1340 (59%)	949 (41%)	2289
Target-dependent [9]	Tar	Celebridades	1734 (50%)	1733 (50%)	3467
Vader [14]	Vad/VAD	Genérico	2897 (69%)	1299 (31%)	4196
SemEval13[5]	S13	Genérico	3183 (73%)	1195 (27%)	4378
SemEval17 [32]	S17	Genérico	2375 (37%)	3972 (63%)	6347
SemEval16 [25]	S16	Genérico	8893 (73%)	3323 (27%)	12216

Tabela 2.1: Domínios e quantidades de instâncias das bases de dados.

palavra diz respeito a um aspecto dela, de forma que palavras que ocorrem em um mesmo contexto são posicionadas em pontos próximos no espaço vetorial [1].

Uma abordagem proposta para o treinamento destas representações foi apresentada em [23]: o *word2vec* (*w2v*). Neste trabalho, são propostas duas novas arquiteturas para o aprendizado de representações distribuídas de palavras: o *Continuous Bag-of-Words* (CBoW) e o *Skip-gram*. Ambas as arquiteturas são redes neurais compostas de três camadas (entrada, projeção e saída) e diferem no fato de o CBoW ser utilizado para predizer uma palavra dado o seu contexto e o *Skip-gram* ser utilizado para predizer as palavras do entorno de uma palavra dada. Mais especificamente, para a situação em que se usa uma janela de tamanho 2 no entorno da palavra, dada uma sequência de palavras que ocorrem juntas *w(t-2)*, *w(t-1)*, *w(t)*, *w(t+1)* e *w(t+2)*, o CBoW prediz *w(t)*, dadas *w(t-2)*, *w(t-1)*, *w(t+1)* e *w(t+2)*, ao passo que o *Skip-gram* prediz *w(t-2)*, *w(t-1)*, *w(t+1)* e *w(t+2)*, dada *w(t)*.

A arquitetura *Skip-gram* foi utilizada em [5] para gerar um modelo estático de *word embeddings* de 400 dimensões, treinado em um conjunto de mais de dez milhões de tweets oriundos do Edinburgh Twitter Corpus [29]. Devido ao seu bom desempenho em análise de sentimentos em tweets [6], este foi o modelo pré-treinado adotado nesta dissertação. As representações das instâncias (tweets) foram computadas como sendo as médias dos *word embeddings* dos tokens obtidos após as etapas apresentadas na Figura 2.1. Um exemplo de

Figura 2.1: Etapas de pré-processamento dos tweets, do tweet original à tokenização.

cálculo dos atributos das instâncias pode ser visto na Figura 2.2, onde os valores utilizados são ilustrativos, não necessariamente condizendo com os valores do modelo.

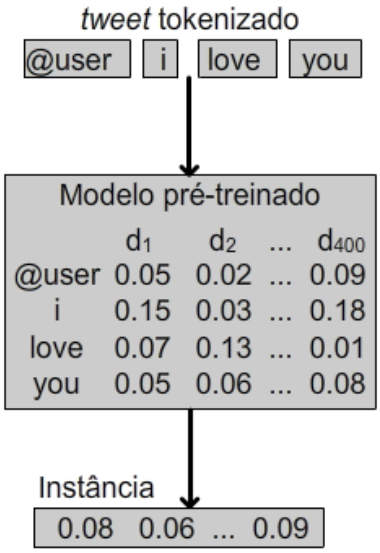


Figura 2.2: Exemplo de cálculo dos atributos com o uso do modelo pré-treinado de *word embeddings*.

2.2.3 Métricas de desempenho

Nos experimentos conduzidos nesta dissertação, foram utilizadas duas métricas de desempenho: acurácia e *F-measure* ponderada (F_1). A acurácia é definida como a quantidade de instâncias do conjunto de teste que foram classificadas corretamente dividida pela quantidade total de instâncias do conjunto de teste.

O cálculo da F_1 é realizado da seguinte maneira:

$$F_1 = \frac{n_p \times F_{1p} + n_n \times F_{1n}}{n_p + n_n} \tag{2.1}$$

Onde n_p e n_n são, respectivamente, a quantidade de instâncias positivas e a quantidade de instâncias negativas do conjunto de teste.

F_{1p} é calculada como:

$$F_{1p} = \frac{2 \times prec_p \times rec_p}{prec_p + rec_p} \tag{2.2}$$

em que:

$$prec_p = \frac{TP}{TP + FP} \tag{2.3}$$

e

$$rec_p = \frac{TP}{TP + FN} \tag{2.4}$$

E F_{1n} é calculada como:

$$F_{1n} = \frac{2 \times prec_n \times rec_n}{prec_n + rec_n} \tag{2.5}$$

em que:

$$prec_n = \frac{TN}{TN + FN} \tag{2.6}$$

e

$$rec_n = \frac{TN}{TN + FP} \tag{2.7}$$

Sendo TP (True Positive), TN (True Negative), FP (False Positive) e FN (Falso Negative) as quantidades de elementos apresentadas nas células da matriz de confusão do conjunto de teste, conforme representado na Tabela 2.2:

		Classe real	
		Positiva	Negativa
Classe predita	Positiva	TP	FP
	Negativa	FN	TN

Tabela 2.2: Matriz de confusão do conjunto de teste.

2.3 Transfer learning for Twitter sentiment analysis: Choosing an effective source dataset

Nesta seção será apresentado um breve resumo das contribuições do artigo do Apêndice A – Transfer learning for Twitter sentiment analysis: Choosing an effective source dataset – publicado no simpósio KDMiLe 2020. Os experimentos conduzidos nesse artigo tinham o objetivo de, dado um conjunto de bases-fonte, verificar se uma combinação de métrica de similaridade e tipo de representação de dados poderia ser utilizada para selecionar

uma base-fonte para treinar um classificador com bom desempenho preditivo para uma base-alvo não-rotulada no contexto de análise de sentimentos em tweets.

Na Subseção 2.3.1 são apresentadas as métricas de similaridade que foram utilizadas como critério para a seleção das bases-fonte, assim como na Subseção 2.3.2 são detalhados os tipos de representação de dados escolhidos para o desenvolvimento do trabalho. O procedimento experimental utilizado para esse artigo está apresentado na Subseção 2.3.3, assim como os resultados obtidos estão apresentados na Subseção 2.3.4.

2.3.1 Métricas de similaridade

Para os experimentos relatados nesse artigo, foram adotadas quatro métricas de similaridade para a seleção da base-fonte mais similar/próxima a cada base-alvo: similaridade de cosseno, distância euclidiana, distância de Jaccard e *Relaxed Word Moving Distance* (RWMD) [16].

2.3.2 Tipos de representações de bases

Para os cálculos das métricas de similaridade, foram adotados cinco tipos de representações para as bases de dados. Cada um destes tipos foi utilizado com um conjunto específico de métricas, como pode ser visto na Tabela 2.3.

Representação de dados	Métricas
Média dos *embeddings* dos tokens	Similaridade de cosseno e distância euclidiana
Média dos *embeddings* das instâncias	Similaridade de cosseno e distância euclidiana
BoW com TF-IDF	Similaridade de cosseno
Conjunto dos tokens lematizados	Distância de Jaccard
Conjunto dos *embeddings* das instâncias	RWMD

Tabela 2.3: Representações das bases de dados e suas respectivas métricas.

Desta forma, a representação das bases como sendo as médias dos *embeddings* dos seus tokens e a representação das bases como sendo as médias dos *embeddings* de suas instâncias foram adotadas para a aplicação das métricas similaridade de cosseno e distância euclidiana. Para a representação das bases usando BoW com TF-IDF foi adotada a similaridade de cosseno, assim como as bases sendo representadas como o conjunto de seus tokens lematizados foi adotada com a distância de Jaccard. A métrica RWMD foi usada com as bases sendo representadas como o conjunto dos *embeddings* de suas instâncias. Todas as métricas foram calculadas para todos os pares de bases de dados. Cabe ressaltar que, a despeito dessas formas de representação das bases para o cálculo das métricas

de similaridade, para o treinamento dos classificadores os atributos das instâncias foram sempre calculados como sendo a média dos *embeddings* de seus tokens, conforme descrito na Subseção 2.2.2.

2.3.3 Procedimento experimental

Para esse artigo foram desenvolvidos dois experimentos. Para ambos o algoritmo escolhido foi o de regressão logística, na implementação do *scikit-learn* [28], com o parâmetro de número máximo de iterações configurado para 10.000 ($max\_iter = 10.000$), com a finalidade de evitar falhas de convergência no algoritmo.

No primeiro experimento, cada uma das 22 bases disponíveis era considerada como base-alvo não-rotulada e as métricas de similaridade entre essa base-alvo e todas as outras 21 bases, consideradas como bases-fonte, eram calculadas. Selecionava-se, então, a base-fonte com maior similaridade à base-alvo e um classificador era treinado com essa base-fonte e aplicado à base-alvo. As métricas de desempenho apresentadas na Subseção 2.2.3 eram calculadas e o seu valor era comparado com o obtido pelas métricas calculadas quando se aplicava um 10-*fold cross-validation* usando a própria base-alvo como conjunto de treinamento. A comparação foi feita calculando-se o ganho, definido como a divisão do valor obtido com o modelo gerado pela base-fonte pelo valor obtido com o modelo gerado pela base-alvo. Embora a base-alvo seja considerada não-rotulada neste experimento, seus rótulos são utilizados para que sejam estabelecidos os valores dos *baselines*. Repetido esse processo para as 22 bases, foi calculado o valor médio de ganho para cada combinação de métrica de similaridade e representação.

No segundo experimento, cada base foi considerada como base-fonte e as 21 bases restantes foram consideradas como bases-alvo. Para cada base-fonte foi treinado um classificador, que era então aplicado a todas as bases-alvo individualmente. Foi feito o cálculo das métricas de desempenho e os resultados obtidos foram comparados com os resultados que são obtidos quando o classificador é treinado com a base-alvo. Foram calculados os ganhos médios dos classificadores treinados em cada uma das bases-fonte.

2.3.4 Resultados

Os resultados do experimento de seleção de bases por meio de uma combinação de critério de similaridade e de representação das bases de dados mostraram que usar a distância euclidiana com a representação das bases sendo calculadas como a média dos *embeddings*

dos seus tokens obteve o melhor resultado preditivo em termos de ganho médio de acurácia, com um desempenho levemente superior a usar a similaridade de cosseno com BoW e TF-IDF. Porém, esta última combinação apresentou um desvio-padrão menor em comparação com a primeira. No que diz respeito aos desempenhos considerando os ganhos médios de F_1, a similaridade de cosseno com BoW e TF-IDF apresentou os melhores resultados, com o menor desvio-padrão. Estes resultados apontam que utilizar essa combinação de métrica e representação dos dados para selecionar uma base-fonte de um conjunto de bases apresenta resultados mais consistentes na capacidade de predizer as classes de uma base-alvo.

Os resultados do segundo experimento, que treinou classificadores com todas as bases-fonte e os aplicou a todas as bases-alvo, mostraram que algumas dessas bases geraram modelos de classificação que, em termos de ganho médio de acurácia, se aproximaram dos valores estabelecidos como *baselines*, ou seja, tiveram desempenho próximo ao obtido quando se utiliza a própria base-alvo como conjunto de treinamento. Considerando os resultados de ganho médio de F_1, essas bases suplantaram os valores obtidos com os classificadores treinados com as bases-alvo, o que mostra que essas bases-fonte possuem uma boa capacidade de generalização. As bases com melhor desempenho neste experimento estão entre aquelas com maior quantidade de instâncias.

2.4 Exploring model transfer strategies for sentiment analysis in Twitter

Nessa seção será apresentado um breve resumo do artigo do Apêndice B – Exploring model transfer strategies for sentiment analysis in Twitter – submetido ao ENIAC 2021. Nesse artigo, partindo-se de possibilidades não cobertas pelo artigo do Apêndice A, foram realizados experimentos com o objetivo de responder a duas questões: **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 [12], 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?*

Na Subseção 2.4.1 estão descritos de forma sucinta os procedimentos experimentais adotados no artigo e, na Subseção 2.4.2, são apresentados os principais resultados obtidos.

2.4.1 Procedimento experimental

Para cada questão a ser respondida pelo artigo, foi realizado um experimento: o Experimento I com a finalidade de responder à questão **Q1** e o Experimento II, com duas estratégias, para responder à questão **Q2**. Em ambos, foi adotado como algoritmo de classificação a regressão logística, na implementação do *scikit-learn*. Novamente, o parâmetro de número máximo de iterações foi configurado para 10.000 ($max\_iter = 10.000$), de forma a evitar falhas de convergência no algoritmo.

Experimento I: o procedimento adotado para este experimento foi, para cada base-alvo do conjunto de 22 bases utilizadas nesta pesquisa, treinar um classificador com as 21 bases restantes, chamada de união das bases-fonte. Esse classificador foi aplicado à base-alvo, calculando-se a acurácia e o F_1 do modelo. Estas métricas foram comparadas ao resultado obtido quando se treina o classificador com a base-alvo usando um 10-*fold cross validation*, por intermédio do ganho de cada métrica, considerado como a divisão do valor obtido com a métrica quando se usa a união das bases-fonte como conjunto de treinamento pelo valor obtido quando se treina o classificador com a base-alvo. Mais uma vez, embora as bases-alvo sejam consideradas não-rotuladas para este experimento, o fato de seus rótulos serem conhecidos permite que se calculem os valores de *baseline* com a própria base-alvo. Além disso, foi feita a comparação da estratégia da união das bases-fonte com a melhor combinação de métrica de similaridade e representação de dados apontada pelo artigo do Apêndice A.

Experimento II: para este experimento foram adotadas estratégias de seleção de instâncias do conjunto união das bases-fonte.

Estratégia S1: nesta estratégia, o conjunto de treinamento foi formado a partir da seleção de instâncias da união das bases-fonte de acordo com uma métrica de similaridade calculada em relação ao centroide da base-alvo. A métrica adotada foi a similaridade de cosseno e o centroide da base-alvo poderia ser calculado de duas maneiras: como a média dos *embeddings* dos tokens dessa base (*tm*) ou como a média dos *embeddings* das instâncias dessa base (*im*).

No que diz respeito à distribuição de classes do conjunto de treinamento, foram consideradas a hipótese de a seleção ser feita mantendo a distribuição original da união das bases-fonte (*ori*) ou gerando um conjunto balanceado (*bal*).

Com a finalidade de se verificar o efeito da diversidade no conjunto de treinamento, foram utilizados dois critérios de seleção: selecionar apenas as instâncias mais simila-

res ao centroide da base-alvo (*sim*) ou selecionar as instâncias mais similares e as mais dissimilares (*dis*), sempre considerando a métrica de similaridade utilizada.

Desta forma, foram testadas oito possibilidades diferentes de configuração e, para cada uma delas, foram selecionados percentuais da união das bases-fonte. Mais especificamente, para cada configuração foram selecionados $p\%, 1 \leq p \leq 100, p \in \mathbb{N}$ da união das bases-fonte. Para cada conjunto de treinamento selecionado foi gerado um classificador, que foi aplicado à base-alvo, sendo calculadas as métricas de desempenho. Os resultados obtidos foram comparados, por intermédio do ganho, aos resultados que se obtêm quando um classificador é treinado com a própria base-alvo usando um 10-*fold cross validation*. Novamente, esta base-alvo é vista como não-rotulada para efeitos da metodologia experimental, mas seus rótulos são utilizados para o estabelecimento dos *baselines*. Os melhores resultados de cada uma das oito configurações também foram comparados aos resultados obtidos com o **Experimento I**.

Estratégia S2: nesta estratégia, os conjuntos de treinamento foram formados a partir da seleção de instâncias da união das bases-fonte de acordo com suas similaridades em relação às instâncias da base-alvo, usando como métrica a similaridade de cosseno.

Os conjuntos de treinamento foram compostos pelas k instâncias da união das bases-fonte que fossem mais similares a cada uma das instâncias da base-alvo, usando como métrica de similaridade a similaridade de cosseno. Iterativamente, para cada instância da base-alvo foi selecionada a instância mais similar na base-fonte, sendo esta instância retirada da lista de instâncias elegíveis à seleção, até que se atingisse o valor desejado de k instâncias para cada instância da base-alvo. O experimento realizado utilizou como critério balancear o conjunto de treinamento. Nesse caso, o número de instâncias da base-fonte de cada classe a serem selecionadas, n_{sel}, para um determinado k, é dado por:

$$n_{sel} = \frac{k \times n_{alvo}}{2} \tag{2.8}$$

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:

$$k \leq \frac{2 \times n_{min}}{n_{alvo}} \tag{2.9}$$

Logo, o valor máximo para k, k_{max}, é definido por:

$$k_{max} = \left\lfloor \frac{2 \times n_{min}}{n_{alvo}} \right\rfloor \tag{2.10}$$

Adicionalmente a essa limitação do valor de k decorrente do fato de o conjunto de treinamento ser balanceado, considerou-se razoável limitar o valor de k de modo que não excedesse 20 instâncias selecionadas por instância da base-alvo. Assim $k \leq min(k_{max}, 20)$.

Para cada conjunto de treinamento foram gerados classificadores, que foram aplicados à base-alvo, sendo calculadas as métricas de desempenho (acurácia e F_1). Os resultados obtidos foram comparados, por meio do cálculo do ganho, com o desempenho de um classificador treinado aplicando um 10-*fold cross-validation* à base-alvo. Os resultados desta estratégia também foram comparados com os obtidos no **Experimento I**.

2.4.2 Resultados

Os experimentos realizados para o artigo do Apêndice B buscavam responder a duas questões. Para a questão **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 [12], 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?), apresentada anteriormente,* foi realizado um experimento que mostrou que utilizar a união de um conjunto de bases-fonte como conjunto de treinamento de um classificador a ser aplicado a uma base-alvo não-rotulada é uma estratégia melhor do que selecionar apenas uma das bases-fonte segundo o melhor critério de similaridade encontrado pelos experimentos desenvolvidos no artigo do Apêndice A. Para apenas duas das 22 bases-alvo, tanto para a acurácia quanto para o F_1, o desempenho do classificador gerado aplicando a abordagem de seleção de uma base-fonte foi superior ao desempenho do classificador treinado com a união das bases-fonte. O ganho médio da estratégia de usar as 21 bases-fonte como treinamento foi superior a 1 o que indica que, na média, o desempenho do classificador gerado com essa abordagem também é superior ao do classificador treinado com a base-alvo.

Para responder à questão **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?*), previamente apresentada, foram desenvolvidas duas estratégias. Na *Estratégia S1*, os resultados apontaram que formar o conjunto de treinamento por meio de uma seleção percentual da união de bases-fonte, levando em

consideração a similaridade das instâncias das bases-fonte em relação à representação de toda a base-alvo, usando como métrica a similaridade de cosseno, consegue aumentar o valor do ganho médio em relação à situação na qual se usa como conjunto de treinamento a união das bases-fonte. Mais precisamente, em seis das oito configurações testadas houve percentuais para os quais o ganho médio superou o ganho de utilizar todas as bases-fonte no treinamento. No entanto, cabe ressaltar que, em todos os casos em que houve esse aumento, o ganho foi modesto e a quantidade de instâncias que teve de ser utilizada para treinar esses classificadores foi bastante elevada, o que mostra que o custo computacional envolvido na seleção não é compensado. Ainda assim, cabe destacar que o melhor desempenho entre as oito configurações ocorreu quando havia um balanceamento do conjunto de treinamento e eram selecionadas as instâncias mais similares e as menos similares à base-alvo.

Para a *Estratégia S2*, os resultados mostraram uma tendência semelhante à apontada pela *Estratégia S1*, ou seja, que utilizar mais instâncias da união das bases-fonte como conjunto de treinamento tende a gerar classificadores com maior poder preditivo. Para 12 das 22 bases de dados houve algum valor de k para o qual o ganho obtido com o classificador gerado a partir de um conjunto de treinamento selecionado por esta estratégia foi maior do que o ganho obtido pelo modelo treinado com toda a união de bases-fonte (**Experimento I**), tanto em termos de acurácia quanto de F_1. Embora isso pudesse indicar que a *Estratégia S2* apresenta um melhor desempenho em relação ao **Experimento I**, a diferença de ganho foi pequena, o que mais uma vez leva à conclusão que o esforço computacional despendido para a seleção de instâncias não é compensado por aumentos consideráveis de ganho. É importante frisar que entre as bases-alvo que tiveram seu valor de k limitado por 20, e não pelo valor de k_{max}, algumas obtiveram o melhor resultado para $k = 20$, o que significa que um aumento no valor de k poderia trazer um aumento no valor de ganho obtido.

2.5 Enriching datasets for sentiment analysis in tweets with instance selection

Nesta seção será apresentado um breve resumo do artigo do Apêndice C – Enriching datasets for sentiment analysis in tweets with instance selection – submetido ao simpósio KDMiLe 2021. Nesse artigo estão relatados experimentos computacionais que foram realizados com o objetivo de investigar abordagens de seleção de dados de um conjunto de bases-fonte rotuladas provenientes de diversos domínios, com o objetivo de enriquecer o

conjunto de treinamento para detecção de polaridade em uma base-alvo também rotulada. Ao contrário dos experimentos relatados nos artigos anteriores, neste artigo considera-se que os rótulos da base-alvo são conhecidos.

Na Subseção 2.5.1 é apresentado o procedimento experimental utilizado para se chegar ao objetivo do artigo, ao passo que na Subseção 2.5.2 são apresentadas as principais conclusões dos experimentos realizados.

2.5.1 Procedimento experimental

Para se atingir o objetivo pretendido com esse artigo foram realizados dois experimentos. Em ambos, o algoritmo de classificação adotado foi o *Support Vector Machines* (SVM), em sua implementação do scikit-learn [28], com o parâmetro de ponderação de classe configurado para a forma balanceada (*class_ weight='balanced'*). Esta nova escolha de algoritmo foi adotada devido ao bom desempenho do SVM para análise de sentimentos em tweets [2]. Como *baselines* foram adotados os valores de acurácia e de F_1 obtidos com o classificador treinado aplicando um 10-*fold cross-validation* à base-alvo.

O primeiro experimento realizado utilizou como abordagem de seleção de dados agregar a maior quantidade possível de instâncias da união das bases-fonte para enriquecer o conjunto de treinamento para a base-alvo, assegurando o balanceamento do conjunto de treinamento. Neste experimento, a base-alvo foi dividida nas mesmas 10 partições utilizadas para a geração do *baseline* e, a cada iteração da validação cruzada, o conjunto de treinamento era formado pelas nove partições da base-alvo e pelos dados selecionados da união das bases-fonte.Um classificador era, então, treinado e aplicado à partição de teste da base-alvo. Para o cálculo do desempenho do classificador nesta situação, foram feitas as médias dos resultados de acurácia e de F_1 das 10 iterações e os valores obtidos foram comparados com os *baselines* por meio do ganho.

No segundo experimento foi adotado um procedimento semelhante ao do experimento anterior, com a base-alvo sendo novamente dividida nas mesmas 10 partições para a execução da validação cruzada. Para cada iteração da validação cruzada é calculada a quantidade mínima de instâncias da classe minoritária que devem ser selecionadas da união das bases-fonte de forma a balancear o conjunto de treinamento. Posteriormente, levando em consideração a quantidade de instâncias restantes na união de bases-fonte é calculada a quantidade de instâncias que corresponda a um percentual especificado dessa união de bases. São, finalmente, selecionadas instâncias da união das bases-fonte considerando as duas quantidades calculadas, segundo a abordagem escolhida. Neste experimento foram

utilizadas três abordagens para a seleção de dados: (I) seleção aleatória das instâncias da união das bases-fonte, (II) seleção das instâncias da união das bases-fonte que sejam mais próximas às instâncias das partições de treinamento da base-alvo segundo o critério de similaridade de cosseno e (III) seleção das instâncias da união das bases-fonte que sejam mais próximas e mais distantes das partições de treinamento da base-alvo segundo o critério de similaridade de cosseno. Para as estratégias II e III, simultaneamente ao critério de similaridade foi adotado o critério de classe, de modo que as instâncias mais próximas ou mais distantes eram selecionadas desde que pertencessem à mesma classe da instância da base-alvo. Para cada uma das iterações foi treinado um classificador com o conjunto de treinamento formado pelas partições da base-alvo e pelas instâncias selecionadas da união das bases-fonte. Este classificador foi aplicado à partição de teste, sendo calculados acurácia e F_1. As médias dos resultados obtidos pelas 10 iterações foram calculadas e serviram como desempenho do classificador para a abordagem e percentual específicos. Os desempenhos foram, então, comparados com os *baselines* por meio do ganho.

2.5.2 Resultados

Os dois experimentos realizados para o artigo do Apêndice C buscavam investigar abordagens de seleção de dados da união de 22 bases-fonte rotuladas para enriquecer o conjunto de treinamento de uma base-alvo rotulada no contexto de análise de sentimentos em tweets.

Para o primeiro experimento foi adotada a abordagem de agregar a maior quantidade possível de instâncias da união das bases-fonte, mas garantindo o balanceamento do conjunto de treinamento. Os resultados deste experimento mostraram que enriquecer o conjunto de treinamento com instâncias da união das bases-fonte aumentou o poder preditivo dos classificadores gerados para a maioria das bases-alvo em relação aos *baselines*. Mais precisamente, tanto para acurácia quanto para F_1, o ganho foi maior ou igual a 1 para 15 das 22 bases-alvo. Ganhos maiores do que 1 significam que o desempenho superou o *baseline*, enquanto ganhos iguais a 1 significam que o desempenho se igualou ao *baseline*.

No segundo experimento, foram adotadas três abordagens distintas para a seleção de dados. A primeira delas selecionou instâncias aleatoriamente de forma a agregar percentuais específicos da união das bases-fonte, assegurando o balanceamento dos conjuntos de treinamento. Nesta abordagem, percentuais baixos de seleção de instâncias (0,5%,

1,0%, 2,5% e 5,0%) quando agregados à base-alvo geraram classificadores com ganhos maiores ou iguais a 1 para 18 das 22 bases em termos de F_1. Embora o melhor resultado em termos de quantidade de vezes que um percentual produziu o melhor desempenho tenha sido com 100,0%, o melhor ranking médio desta estratégia ocorreu com 5,0%, o que mostra que a seleção de pequenas frações de instâncias para enriquecer o conjunto de treinamento, mesmo que feita de forma aleatória, pode aumentar o desempenho preditivo de classificadores.

A segunda abordagem enriquece o conjunto de treinamento selecionando instâncias da união das bases-fonte que sejam mais próximas a cada instância da base-alvo de acordo com o critério de similaridade de cosseno. Essa seleção é feita de forma a garantir o balanceamento e considera, ainda, que as instâncias selecionadas devem ser da mesma classe da instância da base-alvo que originou a sua seleção. Os resultados mostram que a seleção de um percentual baixo (10%) gerou ganho maior ou igual a 1, em termos de F_1 para a maioria das bases-alvo (19 de 22). O melhor desempenho para esta abordagem, tanto em termos de ranking médio quanto em termos de quantidade de classificadores com o melhor desempenho para uma determinada base-alvo, ocorreu com a seleção de 20,0% das instâncias da união das bases-fonte.

A terceira abordagem, por sua vez, se assemelha à segunda, mas neste caso são selecionadas, além das instâncias da união das bases-fonte que estejam mais próximas, também as que estejam mais distantes a cada instância das partições de treinamento da base-alvo. Os resultados apontam que, novamente, percentuais baixos geraram os melhores desempenhos em termos de ganho de F_1. Mais especificamente, selecionar 1,0% ou 2,5% das instâncias produziu ganhos maiores ou iguais a 1 para 19 das 22 bases. No entanto, um percentual intermediário (40,0%) produziu os melhores resultados para cinco bases-alvo e obteve o melhor ranking médio, o que o colocou como o melhor resultado para esta abordagem.

A comparação entre os melhores resultados de cada uma das abordagens utilizadas mostrou que selecionar 40,0% das instâncias da união das bases-fonte, adotando como critério a seleção de instâncias mais próximas e mais distantes, obteve o melhor desempenho. Isso mostra que seleções de percentuais intermediários da união das bases-fonte podem ser suficientes para ampliar, balancear e diversificar o conjunto de treinamento formado por uma base-alvo rotulada de forma a aumentar a sua capacidade preditiva.

Capítulo 3

Conclusões e trabalhos futuros

As pesquisas na área de análise de sentimentos dedicam-se, entre outras tarefas, à detecção de polaridade em textos, classificando-os em positivos ou negativos. Com o advento e a popularização das redes sociais, como Twitter, a quantidade de dados disponíveis para serem analisados vem crescendo nos últimos anos. O Twitter é um serviço de microblog que se caracteriza por textos curtos e informais, chamados tweets, que se constituem desafiadores para a análise de sentimentos devido às suas características de texto informal. Uma estratégia amplamente adotada para a classificação de polaridade em textos é baseada em aprendizado de máquina. Nesta abordagem, características extraídas de textos rotulados de um determinado domínio são usadas como atributos para o treinamento de classificadores que visam detectar a polaridade de textos desse mesmo domínio. No entanto, nem sempre há dados rotulados suficientes para o treinamento de classificadores com um desempenho adequado.

Nesse contexto, uma abordagem adotada dentro da área de transferência de aprendizado é a utilização de dados rotulados de um domínio-fonte para treinar um classificador para um domínio-alvo. Esta dissertação teve como objetivo investigar abordagens de seleção de dados para transferência de aprendizado no contexto de análise de sentimentos em tweets. Com este intuito, foram desenvolvidos diversos experimentos, que resultaram em três artigos, um deles publicado e os outros dois atualmente em processo de revisão. Esses artigos são colocados como apêndices a esta dissertação e um breve resumo de cada um deles é apresentado no Capítulo 2.

Os experimentos que resultaram no artigo do Apêndice A – Transfer learning for Twitter sentiment analysis: Choosing an effective source dataset – investigaram combinações de métricas de similaridade e representações de dados com o objetivo de selecionar, de um conjunto de possíveis bases-fonte, uma base-fonte capaz de gerar um classificador

com bom desempenho para uma base-alvo não-rotulada. Os resultados dos experimentos mostraram que o uso de similaridade de cosseno com representação de bases usando BoW com TF-IDF obteve o melhor desempenho para selecionar uma base-fonte adequada. Adicionalmente, observou-se que algumas bases-fonte possuíam boa capacidade de generalização, independentemente de métricas de similaridade, e que essas bases estavam entre as maiores disponíveis.

O artigo do Apêndice B – Exploring model transfer strategies for sentiment analysis in Twitter – tinha como objetivo investigar estratégias de seleção de dados de bases-fonte rotuladas para serem utilizados como conjunto de treinamento para classificadores para uma base-alvo não-rotulada. Os experimentos apresentados neste artigo mostraram que usar a união de todas as bases-fonte disponíveis como conjunto de treinamento gera classificadores com um desempenho melhor do que o obtido quando o treinamento é feito com a base-fonte mais próxima, segundo a combinação de métrica de similaridade e representação de dados estabelecida no artigo do Apêndice A. Adicionalmente, foram testadas diferentes configurações de seleção de instâncias dessa união de bases-fonte em dois experimentos.

No primeiro, foram selecionados subconjuntos da união das bases-fonte de acordo com critérios de similaridade das instâncias dessa união de bases em relação à base-alvo. Neste experimento foram variadas a forma de representação das bases de dados, o critério de balanceamento do conjunto de treinamento e o critério de similaridade. Os resultados apontaram que exite aumento do desempenho dos classificadores gerados em comparação com o caso em que os classificadores são treinados utilizando a união de todas as bases-fonte, mas que estes resultados são apenas levemente superiores e ocorrem quando são utilizadas quantidades de instâncias muito parecidas com o total de instâncias da união. Cabe ressaltar que os melhores resultados foram obtidos quando o conjunto de treinamento era balanceado. No segundo experimento foi investigada a abordagem de se considerar a similaridade de cada instância da união das bases-fonte a cada instância da base-alvo. Foram feitas seleções incrementais de instâncias para serem utilizadas como conjuntos de treinamento e os resultados deste experimento apontaram que o aumento no tamanho do conjunto de treinamento tende a gerar classificadores com maior poder preditivo.

Os experimentos conduzidos na investigação de estratégias para seleção de dados a serem utilizados no enriquecimento do conjunto de treinamento para gerar classificadores para uma base-alvo rotulada levaram ao artigo do Apêndice C – Enriching datasets for sentiment analysis in tweets with instance selection. Neste artigo, inicialmente é agregada

ao conjunto de treinamento a maior quantidade possível de instâncias da união das bases-fonte de modo que seja garantido o balanceamento deste conjunto. Posteriormente, são adotadas três abordagens para a seleção percentual de dados da união das bases-fonte: seleção aleatória, seleção das instâncias mais próximas e seleção das instâncias mais próximas e das mais distantes. Em todas as abordagens é garantido o balanceamento do conjunto de treinamento.

Os resultados indicam que incrementar o conjunto de treinamento com a maior quantidade possível de instâncias da união das bases-fonte gera classificadores com melhor desempenho preditivo, quando comparados com os classificadores gerados apenas com a base-alvo, para a maioria das bases. No entanto, aplicar as abordagens propostas melhora o desempenho em relação a esse caso inicial. Quando são agregados percentuais baixos de instâncias aleatórias da união das bases-fonte, o desempenho dos classificadores gerados é maior que o desempenho de usar somente a base-alvo como treinamento para quase todas as bases. A seleção das instâncias da união das bases-fonte mais próximas às instâncias da base-alvo gera bons resultados para percentuais baixos, ao passo que utilizar como critério a seleção das mais próximas e das mais distantes gera os melhores resultados para seleções de percentuais intermediários.

Tomando-se por base os resultados obtidos ao longo de toda a investigação para as técnicas de seleção de dados, podem ser observadas algumas conclusões gerais. Quando são utilizadas instâncias de bases-fonte para enriquecer o conjunto de treinamento de classificadores para uma base-alvo rotulada, técnicas de seleção podem gerar um aumento na capacidade preditiva, mas, em geral, esse aumento é baixo e não compensa o esforço computacional envolvido na seleção. No entanto, quando consideramos o caso em que a base-alvo é não-rotulada, selecionar instâncias de um conjunto de bases-fonte para formar o conjunto de treinamento pode gerar classificadores de capacidade preditiva tão boa quanto a de classificadores que pudessem ser treinados com a própria base-alvo. Especificamente para este caso da base-alvo não-rotulada, a utilização da união de bases-fonte de domínios diversos pode ser uma boa alternativa, uma vez que o tamanho do conjunto de treinamento mostrou desempenhar um papel fundamental no desempenho de um classificador. Adicionalmente, pode-se notar que os melhores resultados são obtidos quando há o balanceamento do conjunto de treinamento e quando são considerados critérios de dissimilaridade, que buscam trazer diversidade ao conjunto de treinamento.

Trabalhos futuros podem ser desenvolvidos ampliando o tipo de métricas de similaridade utilizadas para a seleção de dados na transferência de aprendizado no contexto

de análise de sentimentos em tweets. Mais precisamente, podem ser abordadas métricas que envolvam características das distribuições de probabilidade das bases-alvo e fonte. Adicionalmente, podem ser adotadas outras formas de representação dos dados. Com relação à diversidade do conjunto de treinamento, podem ser adotados critérios específicos que busquem tornar este conjunto mais diverso, para além do simples uso de métricas de dissimilaridade entre instâncias das bases-fonte e alvo.

Referências

[1] Agrawal, A., An, A., Papagelis, M. Learning emotion-enriched word representations. In *Proceedings of the 27th International Conference on Computational Linguistics* (Santa Fe, New Mexico, USA, agosto de 2018), Association for Computational Linguistics, p. 950–961.

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

[3] Bengio, Y., Ducharme, R., Vincent, P., Janvin, C. A neural probabilistic language model. *J. Mach. Learn. Res. 3*, null (mar 2003), 1137–1155.

[4] Blitzer, J., McDonald, R., Pereira, F. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing* (USA, 2006), EMNLP '06, Association for Computational Linguistics, p. 120–128.

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

[6] Carvalho, J., Plastino, A. On the evaluation and combination of state-of-the-art features in twitter sentiment analysis. *Artificial Intelligence Review 54* (03 2021).

[7] Chen, L., Wang, W., Nagarajan, M., Wang, S., Sheth, A. Extracting diverse sentiment expressions with target-dependent polarity from twitter. In *ICWSM* (2012).

[8] Diakopoulos, N. A., Shamma, D. A. *Characterizing Debate Performance via Aggregated Twitter Sentiment.* ACM, 2010, p. 1195–1198.

[9] Dong, L., Wei, F., Tan, C., Tang, D., Zhou, M., Xu, K. 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)* (junho de 2014), ACL, p. 49–54.

[10] Go, A., Bhayani, R., Huang, L. Twitter sentiment classification using distant supervision. *Processing 150* (01 2009).

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

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

[13] Guo, J., Shah, D., Barzilay, R. Multi-source domain adaptation with mixture of experts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (Brussels, Belgium, outubro de-novembro de 2018), Association for Computational Linguistics, p. 4694–4703.

[14] Hutto, C. J., Gilbert, E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *ICWSM* (2014), E. Adar, P. Resnick, M. D. Choudhury, B. Hogan, and A. H. Oh, Eds., The AAAI Press.

[15] Jiang, J., Zhai, C. Instance weighting for domain adaptation in NLP. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics* (Prague, Czech Republic, junho de 2007), Association for Computational Linguistics, p. 264–271.

[16] Kusner, M., Sun, Y., Kolkin, N., Weinberger, K. From word embeddings to document distances. In *ICML* (Lille, France, 2015), PMLR, p. 957–966.

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

[18] Liao, X., Xue, Y., Carin, L. Logistic regression with an auxiliary data source. In *Proceedings of the 22nd International Conference on Machine Learning* (New York, NY, USA, 2005), ICML '05, Association for Computing Machinery, p. 505–512.

[19] Liu, B. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*, 2 ed. Studies in Natural Language Processing. Cambridge University Press, 2020.

[20] Liu, M., Song, Y., Zou, H., Zhang, T. Reinforced training data selection for domain adaptation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Florence, Italy, julho de 2019), Association for Computational Linguistics, p. 1957–1968.

[21] Lochter, J., Zanetti, R., Reller, D., Almeida, T. Short text opinion detection using ensemble of classifiers and semantic indexing. *Expert Systems with Applications 62* (06 2016).

[22] Martínez-Cámara, E., Martín-Valdivia, M., López, L., Montejo-Ráez, A. Sentiment analysis in twitter. *Natural Language Engineering 20* (01 2014), 1–28.

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

[24] Mohammad, S., Bravo-Marquez, F., Salameh, M., Kiritchenko, S. SemEval-2018 task 1: Affect in tweets. In *Proc. of The 12th International Workshop on Semantic Evaluation* (junho de 2018), ACL, p. 1–17.

[25] Nakov, P., Ritter, A., Rosenthal, S., Sebastiani, F., Stoyanov, V. SemEval-2016 task 4: Sentiment analysis in Twitter. In *Proc. of the 10th International Workshop on Semantic Evaluation (SemEval-2016)* (junho de 2016), ACL, p. 1–18.

[26] Narr, S., Hülfenhaus, M., Albayrak, S. Language-independent twitter sentiment analysis.

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

[28] 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., Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research 12* (2011), 2825–2830.

[29] Petrovic, S., Osborne, M., Lavrenko, V. The edinburgh twitter corpus. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media* (Los Angeles, CA, junho de 2010), Association for Computational Linguistics, p. 25–26.

[30] Plank, B., van Noord, G. Effective measures of domain similarity for parsing. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (Portland, Oregon, USA, junho de 2011), Association for Computational Linguistics, p. 1566–1576.

[31] Remus, R. Domain adaptation using domain similarity- and domain complexity-based instance selection for cross-domain sentiment analysis. In *2012 IEEE 12th International Conference on Data Mining Workshops* (2012), p. 717–723.

[32] Rosenthal, S., Farra, N., Nakov, P. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proc. of the 11th International Workshop on Semantic Evaluation (SemEval-2017)* (agosto de 2017), ACL, p. 502–518.

[33] Ruder, S. *Neural transfer learning for natural language processing.* Tese de Doutorado, NUI Galway, 2019.

[34] Ruder, S., Ghaffari, P., Breslin, J. G. Data selection strategies for multi-domain sentiment analysis. *CoRR abs/1702.02426* (2017).

[35] Ruder, S., 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* (Copenhagen, Denmark, setembro de 2017), Association for Computational Linguistics, p. 372–382.

[36] Saif, H. Semantic sentiment analysis of microblogs.

[37] Salton, G., Wong, A., Yang, C. S. A vector space model for automatic indexing. *Commun. ACM 18*, 11 (novembro de 1975), 613–620.

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

[39] Speriosu, M., Sudan, N., Upadhyay, S., Baldridge, J. Twitter polarity classification with label propagation over lexical links and the follower graph. In *Proc. of the 1st Workshop on Unsupervised Learning in NLP* (2011), EMNLP '11, ACL, p. 53–63.

[40] Sun, S., Shi, H., Wu, Y. A survey of multi-source domain adaptation. *Information Fusion 24* (2015), 84–92.

[41] Thelwall, M., Buckley, K., Paltoglou, G. Sentiment strength detection for the social web. *J. Am. Soc. Inf. Sci. Technol. 63*, 1 (janeiro de 2012), 163–173.

[42] Van Asch, V., Daelemans, W. Using domain similarity for performance estimation. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing* (Uppsala, Sweden, julho de 2010), Association for Computational Linguistics, p. 31–36.

[43] Xia, R., Hu, X., Lu, J., Yang, J., Zong, C. Instance selection and instance weighting for cross-domain sentiment classification via PU learning. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence* (2013), IJCAI '13, AAAI Press, p. 2176–2182.

APÊNDICE A – Guimarães, Eliseu; Carvalho, Jonnathan; Paes, Aline; Plastino, Alexandre. Transfer learning for Twitter sentiment analysis: Choosing an effective source dataset. VIII Symposium on Knowledge Discovery, Mining and Learning (KDMiLe), 161-168, 2020.

Transfer learning for Twitter sentiment analysis: Choosing an effective source dataset

E. Guimarães[1,2], J. Carvalho[3], A. Paes[1], A. Plastino[1]

[1] Universidade Federal Fluminense, Brazil
[2] Marinha do Brasil
eliseuguimaraes@id.uff.br {alinepaes,plastino}@ic.uff.br
[3] Instituto Federal Fluminense, Brazil
joncarv@iff.edu.br

Abstract. Sentiment analysis on social media data can be a challenging task, among other reasons, because labeled data for training is not always available. Transfer learning approaches address this problem by leveraging a labeled source domain to obtain a model for a target domain that is different but related to the source domain. However, the question that arises is how to choose proper source data for training the target classifier, which can be made considering the similarity between source and target data using distance metrics. This article investigates the relation between these distance metrics and the classifiers' performance. For this purpose, we propose to evaluate four metrics combined with distinct dataset representations. Computational experiments, conducted in the Twitter sentiment analysis scenario, showed that the cosine similarity metric combined with bag-of-words normalized with term frequency-inverse document frequency presented the best results in terms of predictive power, outperforming even the classifiers trained with the target dataset in many cases.

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

Keywords: dataset representation, machine learning, metrics, sentiment analysis, supervised learning, transfer learning

1. INTRODUCTION

Sentiment analysis is a suitcase research problem [Cambria et al. 2017] that involves many Natural Language Processing (NLP) tasks, including the polarity classification of opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities and their attributes expressed in written text [Liu 2012]. With the explosion of social media networks, especially Twitter, people are free to express themselves on any topic using a limited number of characters in short messages called tweets. In this scenario, applying sentiment analysis is particularly challenging considering the characteristics of these short informal messages, such as the incorrect use of grammar, the presence of misspelled words, and lack of context [Martínez-Cámara et al. 2014]. Regarding the polarity detection problem, which aims at identifying whether a text conveys a positive or a negative opinion, two main approaches have been adopted in the literature: lexicon-based methods and machine learning strategies.

Lexicon-based methods rely on the prior polarity of words from existing dictionaries, or lexicons. On the other hand, machine learning strategies, which are the focus of this study, extract characteristics from labeled data in a given domain, called features, and train a model to predict the polarity of new data. However, enough labeled data is not always available, either because the target domain is rare or because manually labeling existent data requires much human effort. In that case, transfer learning approaches emerge as a feasible solution by using labeled data from a different but related source domain to train a classifier to the domain of interest, *i.e.*, the target domain [Pan and Yang 2010].

Nevertheless, choosing between all labeled datasets available from different source-related domains remains a key challenge.

In the context of the challenging issue of choosing an appropriate source dataset, this article aims at determining which metric from a set of distinct distance metrics can be used to identify the most appropriate labeled dataset from a source domain to train a classifier via transfer learning. For this purpose, we evaluate four different distance metrics to select a source dataset, combined with distinct approaches for dataset representation.

The conducted computational experiments, conducted in the Twitter sentiment analysis scenario, showed that the cosine similarity metric combined with bag-of-words normalized with term frequency-inverse document frequency presented the best results, in terms of predictive power, outperforming even the classifiers trained with the target dataset in many cases.

The remainder of this article is organized as follows. Section 2 brings some important concepts used in the article, Section 3 shows examples of similar studies in the literature. Section 4 presents the workflow of the experiments carried out in this study, Section 5 displays and evaluates the results obtained with the experiments, and 6 discusses the conclusions and indicates new research directions.

2. BACKGROUND

In this section, we present some definitions for helping in the comprehension of this article.

Transfer learning: Transfer learning allows the domains, tasks, and distributions used in training and testing to be different [Pan and Yang 2010]. Basically, it uses the source domain and a learning task in this domain to improve the learning for a task in the target domain, using the knowledge obtained in the source domain. It is grounded on the idea that appropriating from prior knowledge and learning can be useful and save resources, avoiding starting from the scratch for every new problem when labeled data is rare or not available. Recently, using transfer learning to solve natural language tasks in the presence of limited data has become a very attractive field of research [Ruder 2019; Devlin et al. 2019].

Word embeddings: Word embeddings [Mikolov et al. 2013] is a technique to represent words in low-dimensional real-valued vectors. Such vectors are learned from large corpora of textual data using neural network techniques aimed at capturing the word's meaning. In that case, words that are frequently used in the same context are represented in the same space.

Cosine similarity: Given two vectors $u = (u_1, u_2, ..., u_n)$ and $v = (v_1, v_2, ..., v_n)$ the cosine similarity (CS) between them is defined as follows:

$$CS = \frac{u \cdot v}{\|u\|.\|v\|} \tag{1}$$

where $u \cdot v$ represents the inner product between u and v, and $\|u\|$ and $\|v\|$ represents their norms.

Euclidean distance: The Euclidean Distance (ED) between two vectors $u = (u_1, u_2, ..., u_n)$ and $v = (v_1, v_2, ..., v_n)$ is defined as follows:

$$ED = \sqrt{\sum_{i=1}^{n}(u_i - v_i)^2} \tag{2}$$

Jaccard distance: The Jaccard Distance (JD) between two sets A and B is defined as the complement of the ratio between their intersection size and their union size. Then:

$$JD = 1 - \frac{|A \bigcap B|}{|A \bigcup B|} \tag{3}$$

Relaxed word moving distance [Kusner et al. 2015]: Consider two datasets D_i and D_j whose word embeddings representations have n and m elements, respectively. The Relaxed Word Moving Distance (RWMD) can be defined as follows:

$$\text{RWMD} = max\left(\sum_{a=1}^{n} f_{ia} \times ED^*_{ia}, \sum_{b=1}^{m} f_{jb} \times ED^*_{jb}\right) \tag{4}$$

where f_{ia} represents the word embedding relative frequency for the *a-th* element of D_i and ED^*_{ia} is the Euclidean distance between the *a-th* element and its closest word embedding in D_j. The terms f_{jb} and ED^*_{jb} are analogous. Thereby, RWMD computes the greatest cost of moving from one dataset to another, weighted by the relative frequencies of the word embeddings, considering both of them as possible origins.

3. RELATED WORK

Distinct studies have been presented in the literature aiming at determining an appropriate distance metric to select data in a given domain to train a classifier to a different target domain via transfer learning [Van Asch and Daelemans 2010; Plank and van Noord 2011; Remus 2012; Ruder and Plank 2017; Santos et al. 2019].

[Van Asch and Daelemans 2010] investigated the relationship between the difference of source and target datasets and the accuracy of Part-of-Speech (PoS) tagger. For the difference calculation, the correlations between six distance metrics and the accuracy of the POS tagger were used, and they showed that Rényi divergence had the best performance in predicting the accuracy of the tagger. In [Plank and van Noord 2011], they studied six metrics and two types of feature representations and their performance in helping select data for transfer learning in parsing tasks in English and Dutch. They found that the variational metric using a topic model representation was the best technique.

Differently, when target data is labeled, [Remus 2012] proposed an approach to select instances from the source dataset based on two metrics: domain similarity and domain complexity. These selected instances and the target dataset were used to compose a new source dataset. Domain similarity was considered based on the idea that selecting the most similar instances to the target dataset could aggregate more information to the trained model. In its turn, the difference between the domain complexities of source and target datasets were used to calculate the reduction to be applied in the original source dataset. The idea behind this was that the more different their complexities are, the less the source data would be useful to compose the new source dataset.

Recently, [Ruder and Plank 2017] proposed an approach to learn data selection measures using Bayesian Optimization for three tasks: sentiment analysis, POS tagging, and parsing. For that purpose, they used six distance metrics as features to learn the new measure, considering three types of dataset representations. Furthermore, they took into consideration that diversity could improve the quality of the training model. Thus, for each training instance, they calculated its diversity, believing that some of them are well suited for knowledge acquisition. The results achieved by them outperformed the existing distance metrics.

[Santos et al. 2019] evaluated three distance metrics on sentiment analysis in the domain of the 2018 Brazilian Presidential Elections using social media data, like tweets, in Portuguese. These metrics were used for datasets selection with the purpose to merge them, and they showed that choosing similar datasets helps in achieving better results. Additionally, they showed that selecting dissimilar datasets worsens the results of the classifiers.

This article differs from previous studies because it investigates, in the scenario of sentiment analysis of tweets, the relationship between distance metrics and the performance of the classifiers trained with the datasets selected by these metrics when applied to the target datasets. Also, in order to conduct our experiments, we have used a large set of 22 Twitter datasets in English.

Dataset	Abbreviation	Positive	% positive	Negative	% negative	Total tweets
irony	iro	22	34%	43	66%	65
sarcasm	sar	33	46%	38	54%	71
ntua	ntu	159	57%	119	43%	278
SemEval15-Task11	S15	47	15%	274	85%	321
sentiment140	stm	182	51%	177	49%	359
person	per	312	71%	127	29%	439
hobbit	hob	354	68%	168	32%	522
iphone	iph	371	70%	161	30%	532
movie	mov	460	82%	101	18%	561
sanders	san	570	47%	654	53%	1224
Narr-KDML-2012	Nar	739	60%	488	40%	1227
archeage	arc	724	42%	994	58%	1718
SemEval18	S18	865	47%	994	53%	1859
debate08	deb	710	37%	1196	63%	1906
HCR	HCR	539	28%	1369	72%	1908
STS-gold	STS	632	31%	1402	69%	2034
SentiStrength	SSt	1340	59%	949	41%	2289
Target-dependent	Tar	1734	50%	1733	50%	3467
VADER	VAD	2897	69%	1299	31%	4196
SemEval13	S13	3183	73%	1195	27%	4378
SemEval17-test	S17	2375	37%	3972	63%	6347
SemEval16	S16	8893	73%	3323	27%	12216

Table I. Datasets characteristics.

4. METHODOLOGY

To conduct the investigation proposed in this article, we used a set of 22 datasets[1] of tweets [Carvalho and Plastino 2020]. Table I presents some characteristics of these datasets, namely their abbreviation, number and fraction of positive and negative tweets, and total number of tweets.

We adopted the following preprocessing steps. First, for each tweet in a given dataset, we replaced URLs and user mentions by unique tokens. Then, all characters were lowercased, and the resulting tweet was tokenized. Finally, we used a pretrained embedding model [Bravo-Marquez et al. 2016], trained over ten million tweets from the Edinburgh Twitter corpus [Petrovic et al. 2010] using the Skip-gram method, to generate a representation for each tweet by averaging the embedding values of its tokens. Henceforth this representation is named as tweet embeddings. We adopted this pretrained model regarding its good performance when compared to other models [Carvalho and Plastino 2020].

To determine the similarity between datasets, we measured the distance between them using the metrics presented in Section 2, i.e., Euclidean distance, cosine similarity, Jaccard distance, and Relaxed Word Moving Distance. For the Euclidean distance, we used two types of representation: dataset embeddings as the average of all word embeddings of the dataset (ED1) and dataset embeddings as the average of all tweet embeddings of the dataset (ED2). The cosine similarity was computed using three forms of representation: bag-of-words (BoW) with term frequency-inverse document frequency (TF-IDF) (CS1), dataset embeddings as the average of all word embeddings of the dataset (CS2), and dataset embeddings as the average of all tweet embeddings of the dataset (CS3). For the Jaccard distance (JD), one more preprocessing step was needed: the lemmatization of the tokens. Then, the lemma sets were considered for the calculation. According to RWMD definition, all word embeddings of the datasets were taken into account for its calculation.

We adopted Scikit Learn's [Pedregosa et al. 2011] implementation of Logistic Regression to train the classifiers. This algorithm was chosen by its good performance in sentiment analysis in Twitter scenario [Carvalho and Plastino 2020]. Specifically, we used each dataset to generate a classification model which was then applied to classify the instances of the other 21 datasets.

[1]Datasets are available at this GitHub repository: `https://github.com/joncarv/air-datasets`

In the experimental evaluation, for each target dataset, we used classification accuracy and weighted average F-measure (F_{AVG}) to compare the results achieved by using the classifier trained with the most similar dataset pointed by the metrics and the results achieved by performing a 10-fold cross-validation when the target dataset is used to train the classifier itself.

Additionally, we compared the classification accuracy and F_{AVG} results achieved when applying the classifiers trained with all datasets, one by one, for each target dataset. When source and target datasets were the same dataset, a 10-fold cross-validation was performed. This comparison intended to verify if some dataset can be selected as source dataset independently of its distance to target dataset with a low predictive loss.

5. EXPERIMENTAL RESULTS

To evaluate the effectiveness of the transfer learning approach, first, we performed a 10-fold cross-validation to induce the logistic regression model for each dataset. Table II presents the results of this evaluation in terms of accuracy and weighted F-measure (F_{AVG}) (second and third columns, respectively). Then, for each target dataset (presented in the rows), we conducted an experiment to identify the most similar dataset to it by using one distance metric at a time, to train a classifier and evaluate its predictive performance on the target dataset. Due to space constraints, we present only the results related to the CS1 metric (fourth and fifth columns), which is the one that has achieved the best overall results, as we shall see later. The sixth and seventh columns present the gain (in %) when the most similar dataset is used in the classification instead of the target dataset itself, in terms of accuracy and F_{AVG}, respectively. The results that increased the performance are presented in boldface type. Finally, the Average and St.dev. rows show the total average gain and its standard deviation, respectively.

Dataset	$Accuracy_{10-FCV}$	$F_{AVG-10-FCV}$	$Accuracy_{CS1}$	$F_{AVG-CS1}$	Accuracy ratio	F_{AVG} ratio
irony	0.66	0.53	0.68	0.68	**102.27%**	**129.60%**
sarcasm	0.56	0.43	0.58	0.53	**102.34%**	**123.55%**
ntua	0.81	0.80	0.86	0.86	**106.24%**	**107.85%**
SemEval15-Task11	0.85	0.79	0.70	0.74	82.47%	94.06%
sentiment140	0.81	0.81	0.69	0.67	84.81%	82.49%
person	0.71	0.59	0.73	0.71	**102.88%**	**119.47%**
hobbit	0.68	0.55	0.69	0.64	**101.70%**	**115.85%**
iphone	0.70	0.57	0.71	0.72	**102.42%**	**126.42%**
movie	0.82	0.74	0.81	0.78	98.70%	**105.20%**
sanders	0.76	0.75	0.61	0.57	80.47%	75.63%
Narr-KDML-2012	0.83	0.83	0.66	0.64	79.00%	77.47%
archeage	0.82	0.81	0.58	0.54	70.02%	66.54%
SemEval18	0.77	0.77	0.63	0.59	81.31%	77.35%
debate08	0.76	0.72	0.64	0.65	85.35%	89.81%
HCR	0.72	0.60	0.73	0.62	**101.31%**	**103.84%**
STS-gold	0.78	0.75	0.80	0.80	**102.69%**	**107.34%**
SentiStrength	0.75	0.74	0.71	0.67	94.35%	89.85%
Target-dependent	0.80	0.80	0.66	0.65	82.82%	81.64%
VADER	0.83	0.81	0.81	0.81	98.02%	**100.06%**
SemEval13	0.77	0.71	0.77	0.73	**101.07%**	**102.83%**
SemEval17-test	0.85	0.85	0.62	0.60	72.47%	71.09%
SemEval16	0.82	0.81	0.80	0.77	96.97%	95.57%
				Average	92.26%	97.43%
				St.dev.	11.30%	18.59%

Table II. Classifiers accuracies and F_{AVG} according to target-dataset model and closest CS1 model and its respectives ratios.

The experiment reported in Table II for CS1 was reproduced for all the distance metrics, and averages and standard deviations presented in the last two rows were summarized in Table III. Table III shows averages accuracy ratio and F_{AVG} ratio on second and fourth columns, respectively, and

Metric	Accuracy ratio average	Accuracy ratio st.dev.	F_{AVG} ratio average	F_{AVG} ratio st.dev.
ED1	**92.72**%	14.08%	95.35%	25.56%
ED2	87.45%	19.44%	89.76%	30.92%
CS1	92.26%	**11.30**%	**97.43**%	**18.59**%
CS2	90.70%	13.76%	93.72%	25.25%
CS3	89.62%	14.38%	92.26%	27.44%
JD	87.94%	15.49%	87.28%	18.76%
RWMD	87.82%	12.09%	88.18%	20.12%

Table III. Averages and standard deviations for accuracy and F_{AVG} ratios according to metrics.

their standard deviations on third and fifth columns. The best results are presented in boldface type. As we can observe, CS1 achieved the best overall results in terms of F_{AVG} (97,43%) with the lowest standard deviation value (18,59%), when used to select a source dataset to train a classifier. It means that, on average, the source dataset selected with this metric achieved 97,43% of the F_{AVG} values obtained by classifiers trained with target datasets. In terms of accuracy, although CS1 achieved the second-best overall result (92,26%), its average performance is comparable to the best overall result achieved by ED1 (92,72%). This represents that CS1 achieved an average of 92,26% of the classification accuracy values when selecting the source dataset in comparison with the classifier trained with the target dataset itself. Nevertheless, CS1 presented the lowest standard deviation value (11,30%), which may indicate that it has a more consistent behavior in selecting a dataset to train a good classifier via transfer learning.

Next, in Tables IV and V, we present an "all versus all" comparison in terms of accuracy and F_{AVG}, respectively. Specifically, each cell in Tables IV and V shows the result achieved by applying the classifier trained on some source dataset (represented in the columns) to classify the instances from some target dataset (represented in the rows). The values in the main diagonal, i.e., the values in cells related to the same dataset in both row and column, refer to the the 10-fold cross-validation evaluation results on the dataset itself. For each target dataset, i.e., each row, the best results are presented in boldface type, and the top five results are underlined. Furthermore, "Top 1" and "Top 5" rows show the number of times each source dataset achieved the best and the top five best results, respectively. For each source dataset, the ratios between the results achieved by the classifier trained on it and the classifier trained on the dataset itself were calculated, and the average of those ratios are shown on "AVG % ratio" row.

	iro	sar	ntu	S15	stm	per	hob	iph	mov	san	Nar	arc	S18	deb	HCR	STS	SSt	Tar	VAD	S13	S17	S16
iro	0.66	0.66	0.37	0.66	0.55	0.34	0.37	0.34	0.34	0.71	0.51	0.62	0.68	0.66	0.66	**0.72**	0.62	0.69	0.54	0.55	0.68	0.68
sar	0.54	0.56	0.52	0.54	0.70	0.46	0.46	0.46	0.46	0.61	0.65	0.72	0.70	0.52	0.56	0.61	0.72	0.68	0.58	0.61	**0.77**	0.59
ntu	0.43	0.53	0.81	0.43	0.79	0.57	0.57	0.57	0.57	0.71	0.80	0.75	0.82	0.58	0.67	0.71	**0.86**	0.74	0.82	0.74	0.73	0.71
S15	**0.85**	0.84	0.25	**0.85**	0.63	0.15	0.15	0.15	0.15	0.81	0.42	0.66	0.70	0.84	0.84	0.84	0.49	0.60	0.32	0.37	0.64	0.45
stm	0.49	0.54	0.68	0.49	0.81	0.51	0.51	0.51	0.51	0.72	0.74	0.71	0.77	0.64	0.61	0.66	**0.82**	0.79	0.77	0.69	0.79	0.74
per	0.29	0.32	0.72	0.29	0.72	0.71	0.71	0.71	0.71	0.61	0.74	0.67	0.69	0.40	0.40	0.57	0.73	**0.77**	0.75	0.72	0.74	0.77
hob	0.32	0.38	0.71	0.32	0.67	0.68	0.68	0.68	0.68	0.45	0.70	0.65	0.52	0.36	0.37	0.39	0.70	0.69	**0.72**	0.69	0.70	0.69
iph	0.30	0.40	0.67	0.30	0.65	0.70	0.70	0.70	0.70	0.53	0.72	0.63	0.62	0.42	0.39	0.42	0.71	0.74	0.74	0.73	0.73	**0.75**
mov	0.18	0.23	0.81	0.18	0.75	0.82	0.82	0.82	0.82	0.43	0.78	0.70	0.58	0.24	0.32	0.38	0.76	0.78	**0.84**	0.81	0.76	0.83
san	0.53	0.56	0.50	0.53	0.69	0.47	0.47	0.47	0.47	**0.76**	0.63	0.62	0.75	0.68	0.59	0.65	0.65	0.69	0.61	0.62	0.75	0.65
Nar	0.40	0.49	0.77	0.40	0.81	0.60	0.60	0.60	0.60	0.73	0.83	0.76	0.82	0.55	0.62	0.66	**0.85**	0.84	0.82	0.80	0.84	0.81
arc	0.58	0.62	0.47	0.58	0.71	0.42	0.44	0.42	0.42	0.69	0.59	**0.82**	0.73	0.64	0.64	0.68	0.65	0.67	0.58	0.58	0.78	0.62
S18	0.53	0.58	0.55	0.53	0.69	0.47	0.47	0.47	0.47	0.69	0.67	0.69	**0.77**	0.61	0.61	0.64	0.72	0.75	0.63	0.66	0.75	0.71
deb	0.63	0.64	0.43	0.63	0.66	0.37	0.37	0.38	0.37	0.67	0.64	0.64	0.67	**0.76**	0.64	0.66	0.69	0.69	0.61	0.64	0.67	0.67
HCR	0.72	0.72	0.31	0.72	0.67	0.29	0.38	0.30	0.28	0.73	0.51	0.72	**0.73**	0.65	0.72	0.73	0.66	0.73	0.66	0.65	0.73	0.73
STS	0.69	0.70	0.56	0.69	0.74	0.31	0.31	0.31	0.31	0.79	0.63	0.67	**0.80**	0.73	0.73	0.78	0.71	0.60	0.56	0.54	0.67	0.51
SSt	0.41	0.46	0.67	0.41	0.72	0.59	0.59	0.59	0.59	0.62	0.72	0.65	0.71	0.51	0.53	0.60	**0.75**	0.72	0.71	0.71	0.71	0.71
Tar	0.50	0.51	0.53	0.50	0.65	0.50	0.51	0.50	0.50	0.62	0.66	0.64	0.69	0.53	0.54	0.60	0.70	**0.80**	0.68	0.66	0.76	0.72
VAD	0.31	0.40	0.76	0.31	0.73	0.69	0.70	0.69	0.69	0.59	0.79	0.65	0.66	0.47	0.50	0.56	0.81	0.79	**0.83**	0.80	0.73	0.80
S13	0.27	0.31	0.75	0.27	0.75	0.73	0.73	0.73	0.73	0.54	0.77	0.73	0.67	0.41	0.45	0.48	**0.81**	0.79	0.77	0.77	0.78	0.77
S17	0.63	0.64	0.40	0.63	0.69	0.37	0.41	0.38	0.37	0.77	0.57	0.78	0.81	0.70	0.67	0.73	0.69	**0.85**	0.62	0.66	0.85	0.80
S16	0.27	0.29	0.73	0.27	0.74	0.73	0.73	0.73	0.73	0.48	0.77	0.67	0.60	0.38	0.38	0.42	0.80	0.80	0.79	0.80	0.78	**0.82**
Top 1	1	0	0	1	0	0	0	0	0	1	0	1	3	1	0	1	**5**	3	3	0	1	2
Top 5	1	1	2	1	6	1	1	1	1	5	6	3	12	3	1	3	**12**	**15**	9	3	**14**	13
AVG % ratio	63%	68%	77%	63%	93%	69%	70%	69%	69%	85%	89%	90%	93%	73%	74%	81%	**95%**	**97%**	89%	88%	**98%**	93%

Table IV. Accuracies for models trained with columns datasets applied to target datasets (rows).

In terms of accuracy (Table IV), we can observe that datasets Target-dependent (Tar column) and SemEval17-test (S17 column) achieved the best overall results regarding their use as source datasets

	iro	sar	ntu	S15	stm	per	hob	iph	mov	san	Nar	arc	S18	deb	HCR	STS	SSt	Tar	VAD	S13	S17	S16
iro	0.53	0.53	0.28	0.53	0.57	0.17	0.23	0.17	0.17	0.65	0.51	0.56	0.63	0.59	0.55	0.65	0.63	0.68	0.54	0.56	0.64	**0.68**
sar	0.37	0.43	0.41	0.37	0.70	0.30	0.30	0.30	0.30	0.55	0.63	0.72	0.69	0.37	0.47	0.53	0.71	0.68	0.53	0.56	**0.77**	0.58
ntu	0.26	0.46	0.80	0.26	0.79	0.42	0.42	0.42	0.42	0.70	0.79	0.75	0.82	0.53	0.66	0.70	**0.86**	0.73	0.81	0.71	0.72	0.67
S15	0.79	0.80	0.24	0.79	0.68	0.04	0.04	0.04	0.04	0.78	0.48	0.71	0.74	0.79	**0.80**	0.79	0.55	0.66	0.35	0.42	0.70	0.51
stm	0.33	0.44	0.65	0.33	0.81	0.34	0.36	0.34	0.34	0.71	0.74	0.71	0.76	0.58	0.57	0.61	**0.82**	0.79	0.77	0.67	0.79	0.73
per	0.13	0.20	0.61	0.13	0.69	0.59	0.59	0.59	0.59	0.62	0.69	0.69	0.71	0.37	0.34	0.58	0.71	**0.78**	0.70	0.67	0.75	0.75
hob	0.16	0.29	0.64	0.16	0.67	0.55	0.55	0.55	0.55	0.41	0.67	0.66	0.52	0.23	0.26	0.33	**0.70**	0.68	0.66	0.64	0.69	0.65
iph	0.14	0.32	0.64	0.14	0.66	0.57	0.59	0.57	0.57	0.52	0.72	0.64	0.63	0.36	0.30	0.36	0.72	0.74	0.72	0.73	0.74	**0.75**
mov	0.05	0.15	0.76	0.05	0.76	0.74	0.74	0.74	0.74	0.46	0.77	0.73	0.62	0.18	0.31	0.39	0.77	0.79	**0.81**	0.78	0.78	0.80
san	0.37	0.44	0.37	0.37	0.69	0.30	0.33	0.30	0.30	**0.75**	0.61	0.57	0.75	0.67	0.49	0.60	0.64	0.69	0.57	0.59	0.75	0.63
Nar	0.23	0.43	0.75	0.23	0.81	0.46	0.45	0.45	0.45	0.73	0.83	0.76	0.83	0.50	0.61	0.64	**0.85**	0.84	0.81	0.78	0.83	0.80
arc	0.42	0.51	0.37	0.42	0.71	0.25	0.29	0.25	0.25	0.66	0.57	**0.81**	0.71	0.60	0.56	0.63	0.65	0.67	0.54	0.56	0.78	0.61
S18	0.37	0.48	0.47	0.37	0.69	0.30	0.31	0.30	0.30	0.65	0.66	0.68	**0.77**	0.53	0.53	0.59	0.72	0.75	0.59	0.64	0.75	0.70
deb	0.48	0.52	0.32	0.48	0.65	0.20	0.22	0.23	0.20	0.58	0.64	0.57	0.60	**0.72**	0.52	0.56	0.68	0.65	0.61	0.65	0.60	0.66
HCR	0.60	0.60	0.21	0.60	0.65	0.14	0.34	0.16	0.12	0.63	0.52	0.62	0.64	0.63	0.60	0.63	0.66	0.64	0.67	0.67	0.62	**0.69**
STS	0.56	0.62	0.55	0.56	0.75	0.15	0.15	0.15	0.15	0.77	0.64	0.68	**0.80**	0.66	0.70	0.75	0.72	0.61	0.55	0.52	0.68	0.49
SSt	0.24	0.36	0.61	0.24	0.72	0.43	0.45	0.43	0.43	0.59	0.71	0.65	0.70	0.44	0.47	0.57	**0.74**	0.73	0.67	0.68	0.71	0.69
Tar	0.33	0.37	0.43	0.33	0.64	0.33	0.36	0.34	0.33	0.56	0.65	0.63	0.67	0.42	0.43	0.54	0.70	**0.80**	0.65	0.64	0.75	0.71
VAD	0.15	0.33	0.71	0.15	0.74	0.57	0.59	0.57	0.56	0.59	0.78	0.66	0.67	0.43	0.48	0.55	**0.81**	0.79	0.81	0.78	0.74	0.79
S13	0.12	0.19	0.68	0.12	0.75	0.61	0.62	0.61	0.61	0.54	0.73	0.74	0.68	0.36	0.43	0.47	**0.80**	0.79	0.71	0.71	0.78	0.73
S17	0.48	0.52	0.26	0.48	0.70	0.20	0.29	0.22	0.20	0.75	0.55	0.77	0.80	0.66	0.57	0.69	0.69	**0.85**	0.60	0.65	0.85	0.80
S16	0.12	0.15	0.65	0.12	0.75	0.61	0.62	0.61	0.61	0.47	0.74	0.69	0.61	0.32	0.31	0.38	0.80	**0.81**	0.74	0.77	0.79	0.81
Top 1	0	0	0	0	0	0	0	0	0	1	0	1	2	1	1	0	**7**	**4**	1	0	1	3
Top 5	1	1	1	1	10	0	0	0	0	3	5	4	10	2	1	3	**16**	**16**	5	5	**16**	11
AVG % ratio	47%	59%	74%	47%	101%	54%	58%	55%	54%	89%	95%	98%	100%	70%	71%	81%	**104%**	**105%**	94%	94%	**106%**	99%

Table V. F_{AVG} for models trained with columns datasets applied to target datasets (rows).

in the classification via transfer learning. While the dataset Target-dependent achieved the top five best results in 15 out of the 22 datasets (97% of AVG % ratio), dataset SemEval17-test achieved the top five best results in 14 out of the 22 datasets (98% of AVG % ratio). It is worth mentioning that dataset SentiStrength (SSt column) achieved the best overall results in five out of the 22 datasets, and the top five best results in 12 out of the 22 datasets (95% of AVG % ratio). These ratios indicate the average gain of classification accuracy achieved by the source dataset in one column compared to the classifier's accuracy results trained with the target dataset itself. That means they had almost the same accuracy of the target dataset classifier, which is quite remarkable.

Similarly, in terms of F_{AVG} (Table V), we can notice that datasets SemEval17-test, Target-dependent, and SentiStrength also achieved the best overall results. Their AVG % ratios for the F_{AVG}, respectively 106%, 105%, and 104%, outperformed the results obtained using the classifiers trained with the target datasets themselves. Interestingly, these three datasets are among the ones with the greatest number of tweets, which could indicate why they had such a good performance, independently of the distance to the target dataset. Moreover, the variety in SemEval17-test and Target-dependent subjects, respectively entities, products, and events, and celebrities, products, and companies, may help to explain their performance.

6. CONCLUSIONS AND FUTURE WORK

This article intended to determine the most suitable distance metric between two datasets to choose a labeled dataset to train a target classifier via transfer learning. For this purpose, we evaluated four types of metrics in a large set of 22 Twitter datasets in English, achieving promising results.

In fact, one particular combination of distance metric and dataset representation reached a notorious performance over the seven combinations employed: the cosine similarity applied to the datasets represented with BoW and TF-IDF (CS1). This metric achieved the best results in term of F_{AVG} and the second best in terms of accuracy. In terms of accuracy, the best metric was ED1, although that value was very close to CS1's accuracy. Moreover, the CS1 metric presented the smallest standard deviations, showing that it has a more consistent behavior in predicting the target dataset's classes. This result reveals that selecting CS1 as the distance metric to choose a training dataset tends to reach good results in most of the cases.

Furthermore, the experiment conducted to verify if some dataset, independently of a distance metric, could be selected to build a proper performance classifier revealed that some of the datasets reach good generalization. SemEval17-test, SentiStrength, and Target-dependent had good results in terms of both accuracy and F_{AVG}. On average, they displayed a greater F_{AVG} value than the classifier

trained by the target dataset itself.

Future work could use more distance metrics or change the datasets representation form to establish a closer relationship between a distance metric and performance metrics. Also, identifying which characteristics of those datasets lead to the best performance is a promising path for future investigation. It can start by extracting features from these datasets, like their dimension, or the vocabulary size. In addition, identifying when to rely on the distance metric or when to adopt the dataset with the best overall performance is a promising venue for future work.

ACKNOWLEDGMENT

The authors thank the agencies CNPq and FAPERJ for the financial support.

REFERENCES

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 International Conference on Web Intelligence (WI)*. IEEE, Omaha, USA, pp. 536–539, 2016.

Cambria, E., Poria, S., Gelbukh, A., and Thelwall, M. Sentiment analysis is a big suitcase. *IEEE Intelligent Systems* 32 (6): 74–80, 2017.

Carvalho, J. and Plastino, A. On the combination and evaluation of state-of-the-art features in twitter sentiment analysis. *Artificial Intelligence Review*, 2020.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*. Association for Computational Linguistics, Minneapolis, MN, 2019.

Kusner, M., Sun, Y., Kolkin, N., and Weinberger, K. From word embeddings to document distances. In *Proceedings of the International Conference on Machine Learning*. PMLR, Lille, France, pp. 957–966, 2015.

Liu, B. *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers, San Rafael, USA, 2012.

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.

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

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.

Petrovic, S., Osborne, M., and Lavrenko, V. The edinburgh twitter corpus. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media*. Association for Computational Linguistics, Los Angeles, CA, pp. 25–26, 2010.

Plank, B. and van Noord, G. Effective measures of domain similarity for parsing. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Portland, USA, pp. 1566–1576, 2011.

Remus, R. Domain adaptation using domain similarity- and domain complexity-based instance selection for cross-domain sentiment analysis. In *Proceedings of the 2012 IEEE 12th International Conference on Data Mining Workshops*. IEEE, Brussels, Belgium, pp. 717–723, 2012.

Ruder, S. *Neural transfer learning for natural language processing*. Ph.D. thesis, NUI Galway, 2019.

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*. Association for Computational Linguistics, Copenhagen, Denmark, pp. 372–382, 2017.

Santos, J. S., Paes, A., and Bernardini, F. Combining labeled datasets for sentiment analysis from different domains based on dataset similarity to predict electors sentiment. In *Proceedings of the 2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*. IEEE, Salvador, Brazil, pp. 455–460, 2019.

Van Asch, V. and Daelemans, W. Using Domain Similarity for Performance Estimation. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*. Association for Computational Linguistics, Uppsala, Sweden, pp. 31–36, 2010.

APÊNDICE B – Guimarães, Eliseu; Carvalho, Jonnathan; Paes, Aline; Plastino, Alexandre. Exploring model transfer strategies for sentiment analysis in Twitter. Submetido ao XVIII Encontro Nacional de Inteligência Artificial e Computacional (ENIAC), 2021.

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 computacionais 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.

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

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

Tabela 1. Características das bases de dados de tweets.

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

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 Computacionais

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.

APÊNDICE C – Guimarães, Eliseu; Vianna, Daniela; Paes, Aline; Plastino, Alexandre. Enriching datasets for sentiment analysis in tweets with instance selection. Submetido ao IX Symposium on Knowledge Discovery, Mining and Learning (KDMiLe), 2021.

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

Table 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 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 o procedimento

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

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 esta quantidade, verifica-se qual será a classe minoritária 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, de forma a garantir o balanceamento, 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 e as seleções foram realizadas de forma a garantir o balanceamento. 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. 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.

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**

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

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.

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

Table 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

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

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 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

Table 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 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,

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

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

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**

Table 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.