Cross-Domain Sentiment Analysis in Brazilian Portuguese

Larissa F. S. Britto[1], Raissa Camelo[2]
[1,3]Departamento de Computação (DC) - Universidade Federal Rural de Pernambuco (UFRPE)
Recife, Pernambuco, Brazil
larissa.feliciana@ufrpe.br[1], raissa.camelo@ufrpe.br[2]

***Abstract*—Over the years, many approaches to the cross-domain task in Sentiment Analysis have been proposed. The vast majority of such approaches concerns corpora in English language. Following a different methodology, this paper proposes a new product review dataset written in Brazilian Portuguese. This dataset are freely available and they cannot only be used as benchmark dataset by other researchers, but also to perform domain adaptation comparative assessments. In addition, the entire creation process for the proposed dataset is described. Another contribution resides in the experimental evaluation of the state-of-the-art algorithm Structural Correspondence Learning (SCL) and Bidirectional Encoder Representations from Transformers (BERT) on cross-domain sentiment analysis (domain adaption).**

***Index Terms*—cross-domain·sentiment analysis, structural correspondence learning, domain adaptation**

I. Introduction

The increasing popularity of social networks, blogs, and forums enable more people to expose their opinions on the Internet. This phenomenon generates a huge amount of textual data of great interest by the Text Mining community, especially in Sentiment Analysis (SA), which concerns the computational study of opinions on various subjects, products, and entities [1].

Despite the large number of opinions available online and the recent advances in this area, SA systems still have difficulty in handling opinions present in unstructured textual data, i.e., the precise extraction and categorization of sentiments is still a challenging task. This is even more evidenced in the scenario in which the researcher has to deal with textual data from different domains [2]. For instance, words whose meaning or polarity (positive, negative or neutral) depends on the domain, and discriminative words in one domain that do not appear in another, are some of the problems encountered by researchers when performing sentiment analysis. Dealing with this issue has great importance in SA, since annotated data from many domains are scarce. One approach to deal with this problem is Cross-Domain Sentiment Analysis (CDSA) [2] which attempts to alleviate scarce annotated data by proposing more flexible solutions to domain adaptation. Indeed, many approaches to CDSA have been proposed and the vast majority of them concerns corpora in English language [3] [4] [5].

In this paper, we experimentally evaluate, adopting methodology established in the literature, the performance of the Structural Correspondence Learning algorithm (SCL) [6] and BERT language representation in CDSA on two datasets in Brazilian Portuguese language. The SCL was selected in this work because is one of the first and most discussed domain adaptation algorithms in CDSA. Moreover, to evaluate it performance, a new datasets of product reviews in the domain of games in Brazilian Portuguese are proposed to the SA scientific community.

The main contributions of this work are threefold:

1) The development of one SA datasets written in Portuguese language for SA analysis and, in special, for CDSA;
2) An empirical evaluation using two different domains of products reviews taking into account pivot features;

The remainder of this paper is organized as follows. Section II describes BERT representation. Section III presents an overview on cross-domain sentiment analysis approaches, mainly SCL. The SCL algorithm used in this work are described in Section IV. Section V describes the entire dataset development methodology. . Our experimental setup and results are described in Section VI. Section VII concludes this paper.

II. Bidirectional Encoder Representations from Transformers

Bidirectional Encoder Representations from Transformers (BERT) is a simple and powerful language representation model create to alleviates standard unidirectional models limitations. It is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. This pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks [7]. For this work, we follow the "BERT Text Classification in 3 Lines of Code Using Kera" [1] tutorial implementation witch use keras to load the pre-trained BERT model with a randomly initialized final Dense layer.

III. Cross Domain Sentiment Analysis

CDSA aims to generalize a classifier by transferring the knowledge from a labeled domain, named the source domain, to an unlabeled domain, a.k.a the target domain [8].

This is a challenging task mainly due to the domain dependence issue in Sentiment Analysis [2]. Over the years many

[1]https://towardsdatascience.com/bert-text-classification-in-3-lines-of-code-using-keras-264db7e7a358

domain adaptation algorithms were proposed, using different approaches as Word Embedding [9] [10], Transfer Learning [11] [12], Lexicons or Thesaurus [13] and Deep Learning [14].

One of the first proposed algorithms for CDSA is the Structural Correspondence Learning (SCL) [6] which introduces the concepts of pivot features, (i.e. common relevant features that appear in both source and target domains). The pivot idea has inspired many other algorithms for CDSA, including Spectral Feature Alignment (SFA) [11], which aligns domain-specific features from different domains using Spectral Clustering algorithm. SFA is also based on the knowledge learned from pivot features, or domain independent words.

In [9], an embedding model was proposed for both the training phase of CDSA algorithm and constructing three objective functions that capture: (a) the distributional properties of the pivots, (b) the label constraints in the source domain documents, and (c) the geometric properties in the unlabeled documents in both source and target domains. Better performance can be achieved by optimizing all the above objective functions.

In addition, SCL has also inspired some variants such as [15] in which a weighted SCL model (W-SCL) was proposed. W-SCL algorithm assigns small values as weights to high-frequency domain-specific features, and higher weight values to examples with the same label as the involved pivot feature.

In [16], two SCL variants based on neural networks models were proposed: Autoencoder Structural Correspondence Learning (AE-SCL) and Autoencoder Structural Correspondence Learning with Similarity Regularization (AE-SCL-SR). These models use Autoencoders Neural Networks to learn SCL representation.

IV. Structural Correspondence Learning

Cross-domain Sentiment Analysis has received a lot of attention from researchers with many approaches proposed over the last decade. Among these approaches, Structural Correspondence Learning [6] introduces a method to identify correspondences among features from different domains. The main idea of SCL is to make a correspondence between *non-pivot features*, (terms less significant or that do not occur in one of the domains) and the *pivot features* (meaningful terms important for both domains) by means of calculating their correlation.

The SCL algorithm is basically divided into three stages: In the first step, the pivot features are selected, then a semi-supervised stage aiming to learn the SCL representation (correlation between features from domain A and B) using both labeled data from the source domain (A) and unlabeled data from the target (B) domain. Finally, in the supervised stage, a mapping matrix is applied to both domains to make the correspondence between them. As a result, such new pivot features are used for training a supervised classifier. The underlying working hypothesis in SCL is that if the pivot features have been correctly chosen, the mapping matrix will produce a good feature encoding. In other words, a new subspace of optimized features can be used as input by the classifier. Those steps are better illustrated in Fig. 1.

In the SCL algorithm, pivot features are defined as terms or phrases that occur frequently in both domains (source and target) [6], thus, having high discriminative power. Therefore, their selection is one of the most important steps for achieving good performance in CDSA. Due to the paramount importance of having well selected features as pivot features, the SL algorithm was extended by [17], in which pivot features are selected using Mutual Information defined as the ratio between the probability of observing two variables x and y together and the probabilities of observing x and y independently [18].

Input: labeled source data $\{(\mathbf{x}_t, y_t)_{t=1}^{T}\}$, unlabeled data from both domains $\{\mathbf{x}_j\}$

Output: predictor $f : X \rightarrow Y$

1. Choose m pivot features. Create m binary prediction problems, $p_\ell(\mathbf{x}),\ \ell = 1 \ldots m$

2. For $\ell = 1$ to m

$\hat{\mathbf{w}}_\ell = \underset{\mathbf{w}}{\text{argmin}} \left(\sum_j L(\mathbf{w} \cdot \mathbf{x}_j, p_\ell(\mathbf{x}_j)) + \lambda ||\mathbf{w}||^2 \right)$

end

3. $W = [\hat{\mathbf{w}}_1 | \ldots | \hat{\mathbf{w}}_m]$, $[U\ D\ V^T] = \text{SVD}(W)$, $\theta = U^T_{[1:h,:]}$

4. Return f, a predictor trained on $\left\{ \left(\begin{bmatrix} \mathbf{x}_t \\ \theta \mathbf{x}_i \end{bmatrix}, y_t \right)_{t=1}^{T} \right\}$

Fig. 1. The SCL Algorithm extracted from [6]

A. Feature Correlation

The main goal of the pivot features is to map the original feature space to a *shared feature space* and this is done by calculating the correlation between non-features and features.

In [6] [17] a binary classifier is created for each pivot, the input of these classifiers are the non-pivots of each instance of the unlabeled data, and the output corresponds to the occurrence of the pivot feature in this instance. The matrix of weights obtained by the classifiers after training is post-processed using Singular Value Decomposition (SVD), used to matrix decomposition. In [19], Autoencoder Structural Correspondence Learning (AE-SCL) was proposed, using autoencoder neural networks to learn to encode the non-pivot features representation of each instance into a low dimensional feature space. In the same work, Autoencoder Structural Correspondence Learning with Similarity Regularization (AE-SCL-SR) was also proposed, with the addition of a pre-training step with a word embedding model. After the generation, the matrix is applied as a transformation matrix.

V. Corpora Development

Two datasets in Brazilian Portuguese from different domains were used in this work. Three of them were created especially

for this paper[2], whereas the fourth one is better described in [16]. Both the domain of above datasets and their sources are described next.

Electronics (E) This dataset was obtained from [16]. It is composed of product reviews, most of them are electronics including smartphones and kitchen products, The reviews were gathered from Mercado Livre site.[3]

Games (G) This dataset is formed by game reviews extracted from Steam [4] using its Web API. in Brazil.
The reviews were obtained using AppFollow API [5].

Fig. 2 depicts the pipeline-based processing to build the corpora.

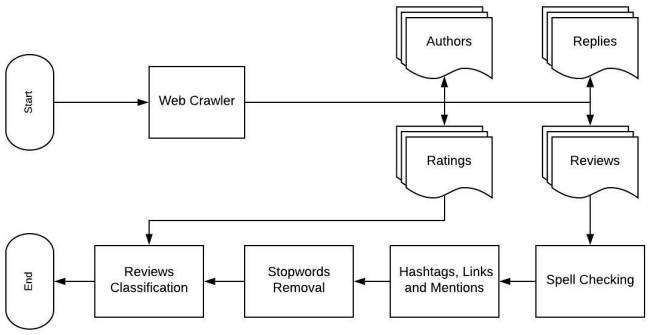

Fig. 2. The full process of dataset creation.

A. Corpora Preprocessing

In the first step the data were obtained using APIs and web scraping, the tool Beautiful Soup[6] was employed. After gathering the reviews containing information about authors, review text, rating and the replies for a given review, the normalization step takes place to provide a standardized version of the input reviews. Usually, online comments have spelling errors, slang, and abbreviations. NLTK[7] and Regular Expression were used to normalize all the reviews.

The remainder of the pipeline steps are:

i) Spell Checking - The most common spelling errors were corrected using a dictionary as well as some internet slang and abbreviations. For example, the word ”não”/”no” is commonly replaced for the letter ’n’. In this step, repeated letters used by the users to emphasize their opinion strength were removed.

ii) Special Terms - In this step, common formations commands used especially on the internet like hashtags, links and mentions are replaced.

iii) Stopwords Removal - Especially for this work punctuation, stopwords and numbers were removed. Some stopwords that could indicate polarity inversion (e.g. ”não”/”no”, ”sem”/”without” and ”mas”/”but”) were not filtered out [20]. All words were converted to lowercase and words with a frequency less than 20 in all the dataset were removed.

To demonstrate the effects of each step, we use the following comment as an input example:

Esperava mais, n gostei! muito chaaaaato #resenha

At the end of the pipeline, the result sentence is:

esperava mais não gostei muito chato tag

B. Corpora Statistics

We used NLTK and Scikit-Lean[8] toolkits to obtain information about the datasets used in this work. Tab. I summarizes some basic statistics: the number of reviews (positives and negatives), the number of words in the corpus, and the average number of words per sentence. Tab. II shows the most relevant words for each dataset (in Portuguese and English), according to their tf-idf weights.

T

VI. Experiments

This section presents the experimental evaluation of the SCL algorithm on the dataset proposed by this work. Fig. 3 displays the main steps of all the experiments reported in this section, including preprocessing, pivot selection, and model generation.

First the unlabeled data from both domains were pre-processed (as described in Corpora Preprocesseing section) and their features extracted, resulting in a binary vector pointing the presence or absence of that feature in the document. Then, the features pivots were selected, and the representation model were trained using the unlabeled pre-processed data. The interesting labeled data from both domain were also pre-processed and have their feature extracted, and the representation matrix applied. Finally, the classification model was generated, trained and tested on the labeled data after the matrix was applied.

A. Experimental Setup

To the best of our knowledge, datasets in Portuguese have never been evaluated before using CDSA algorithms. For this reason, we cannot provide a direct and fair comparative assessment with other previous works. Therefore, in all the experiments, we defined our baseline performance using the Logistic Regression algorithm without any kind of adaptation (NO-ADPT). In other words, we built a classifier using the entire source domain dataset (A) as training data, and then, predicting all the examples from the target domain dataset (B) using this classifier.

Aiming at evaluating the SCL-based algorithms aforementioned, we randomly sampled, from all the two product review datasets, 2000 reviews (1000 positive/1000 negatives). We

[2]https://github.com/larifeliciana/Sentiment-Analysis-Portuguese-Datasets
[3]https://www.mercadolivre.com.br
[4]https://store.steampowered.com/
[5]https://appfollow.io/
[6]https://www.crummy.com/software/BeautifulSoup/bs4/doc/
[7]https://www.nltk.org/book/
[8]https://scikit-learn.org/stable/

TABLE I
DATASETS STATISTICS

domain	#reviews	#positive reviews	#negative reviews	#words	words/sentence av.
games	70,503	52,761 (75%)	17,742 (25%)	1,658,091	80
electronics	43,321	21,819 (50%)	21,5499 (50%)	605,131	49

TABLE II
MOST RELEVANT WORDS ACCORDING TO TF-IDF SCORES

domain	PT	EN
games	bem, bom, game, história, jogar	good, good, game, story, play
electronics	bom, conclusão, contras, excelente, nada	good, conclusion, cons, excellent, nothing

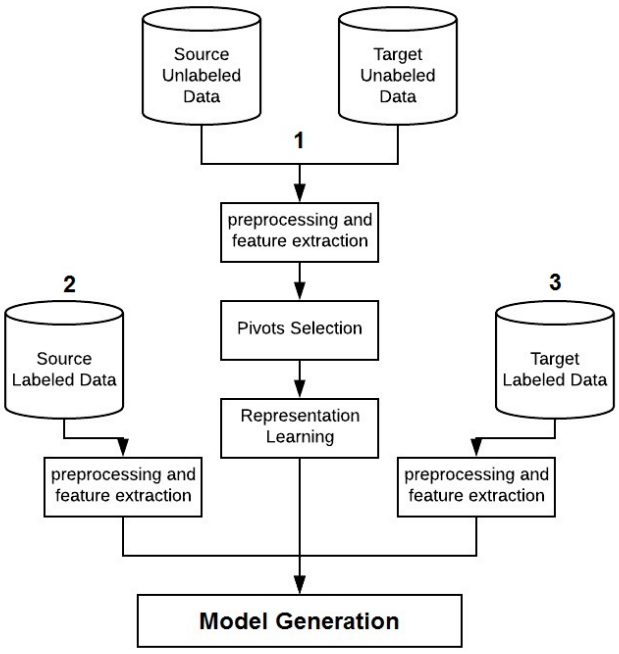

Fig. 3. Main steps of the original SCL algorithm

found the pivot features (pivot selection step) using the following number of randomly sampled reviews of the unlabeled target datasets: 25,305 (E)letronics, and 20,094 (A)pps.

A total of 12 combinations have been acquired, each of them alternating either the source or the target domain. Notice that, when a given dataset is used as a target domain, its label information is ignored.

For comparison purposes, we adopt the test accuracy (defined as the ratio between the total number of correctly predicted labels from both positive and negative classes and the total number of dataset samples) as the evaluation metric,

B. Hyperparameter Tunning

To find the best hyperparameters of the final classification models, we adopted a Grid Search procedure. Besides the 2 possible combinations obtained from the 2 datasets (as source and target domains alternatively), there is also several options concerning the number of pivots and dimensionality, as summarized in Tab. III. The hyperparameters values were inspired on [19]. The following number of hyperparameter combinations were performed for each algorithm: 12 (NO-ADAPT), 144 (SCL-MI). As in [19], we select those hyperparameter values that yielded the best performance in the grid search preliminary experiments.

C. Results and Discussion

This section presents the results and discusses the experiments with the goal of analyzing the impact of using SCL algorithms on the selected datasets. Tab. IV shows the results for all experiments.

As expected, the classification results on different domains without applying any adaptation have the worse scores, presenting lower accuracies in the experiments. On the one hand, a comparison between the SCL-MI showed that the former obtained a positive difference in terms of accuracy. Also, SCL-MI had lower scores in only 3 combinations in comparison to all other approaches (see Tab. IV). Such results are in accordance with [19]. Binary and TF-IDF were the best methods to feature extraction, and SVM were the classifier with the best performance considering accuracy metric.

Our experiments have been performed in different domains in comparison to the works presented by [6], [19], due to the lack of data in Portuguese in such domains. That may be considered a limitation, once the ease or complexity of domain adaptation depends on how closely related or otherwise the source and target domains are to each other [2]. Domains that could be as close as those used in the English language work were used.

To understand how the core stage of pivot selection in SCL algorithm is important, we inspected the top pivot features from one of the best models and one of the worst, as shown in Tab. V. According to the table, the best models were able to select more pivots denoting meaningful phrases related to subjective sentiment opinions than the others. Indeed, although there are some pivots in common in both models, the worst models selected many non-discriminative words, i.e., unigrams as pivot features, such as ”ler”/”read”, ”história”/”history”., whereas the best models selected more meaningful phrases including ”não recomendo”/”I don’t recommend” and ”não gostei”/”i didn’t like it”. Moreover, the neural models also

TABLE III
HYPERPARAMETERS

Hypeparameter	SCL-MI [17]
Number of pivots	200,300,400
Dimentionality	40, 50, 100, 150

TABLE IV
EXPERIMENTS

	TF		TF-IDF		IDF		BINARY	
	N.A	SCL	N.A	SCL	N.A	SCL	N.A	SCL
LR (G/E)	0.8235	0.83	0.8185	0.795	0.8095	0.7915	0.8235	0.83
RF (G/E)	0.7305	0.6655	0.7765	0.6935	0.74	0.722	0.7415	0.6915
SVM (G/E)	0.7755	0.8185	0.8005	**0.8345**	0.7835	0.819	0.7755	0.8185
LR (E/G)	0.681	0.7115	0.6895	0.7035	0.672	0.6965	0.681	0.7115
RF (E/G)	0.6095	0.6125	0.629	0.6605	0.6235	0.646	0.6395	0.617
SVM (E/G)	0.613	0.6785	0.6945	**0.7315**	0.693	0.722	0.613	0.6785

have more pivot features with stronger sentiment polarity, as in the selected pivot phrases ”melhor jogo”/”best game”, ”lixo”/”garbage”, and ”excelente”/”excellent”.

Therefore, as shown above, selecting the best features as pivots in the SCL-based approache, evaluated in this research work, can effectively improve accuracy performance in CDSA, thanks to the better mapping process that learn more similar representations, even when applied in the Brazilian Portuguese language.

Tab. VI shows BERT results.

SCL reachs the best performance with SVM classifier and TFIDF method, achieving accuracy of 73% and 83%, BERT reachs 75% and 86% of accuracy, getting the best performance.

VII. CONCLUSIONS

In this paper, we investigated the cross-domain sentiment analysis problem and, in special, the state-of-the-art algorithm SCL. Thanks to its success in previous work, the SCL was extended over the years, and this work described our main contribution in terms of both the proposed experimental setup and the creation of new datasets to be adopted for future researches. Our experiments showed that among the classification models induced, the clear winner was the models witch uses SCL to the domain adaptation. Another major point concerned the analysis of pivot selection step which retrieved, representative pivot features for the input dataset, with a strong positive impact on the best models. The experiments showed also that uses BERT have improved the accuracy. As future work, we intend to extend the comparative evaluation with other state-of-the-art CDSA algorithms, again considering datasets in Portuguese. Finally, it would be interesting exploring more specific features of Portuguese language in CDSA, thus evaluating the impact of language particularities.

REFERENCES

[1] Z. Zuo, “Sentiment analysis of steam review datasets using naive bayes and decision tree classifier,” 2018.

[2] T. Al-Moslmi, N. Omar, S. Abdullah, and M. Albared, “Approaches to cross-domain sentiment analysis: A systematic literature review,” *IEEE Access*, vol. 5, pp. 16 173–16 192, 2017.

[3] B. Gupta, S. Awasthi, P. Singh, L. Ram, P. Kumar, B. R. Prasad, and S. Agarwal, “Cross domain sentiment analysis using transfer learning,” in *2017 IEEE International Conference on Industrial and Information Systems (ICIIS)*, Dec 2017, pp. 1–5.

[4] O. Abdelwahab and A. Elmaghraby, “Deep learning based vs. markov chain based text generation for cross domain adaptation for sentiment classification,” in *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, July 2018, pp. 252–255.

[5] H. Yu, Y. Pan, and C. Zhou, “Domain adaptation approach for sentiment analysis,” in *Proceedings of the 2019 3rd International Conference on Compute and Data Analysis*, ser. ICCDA 2019. New York, NY, USA: ACM, 2019, pp. 94–97. [Online]. Available: http://doi.acm.org/10.1145/3314545.3314553

[6] J. Blitzer, R. McDonald, and F. Pereira, “Domain adaptation with structural correspondence learning,” in *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP ’06. Stroudsburg, PA, USA: Association for Computational Linguistics, 2006, pp. 120–128. [Online]. Available: http://dl.acm.org/citation.cfm?id=1610075.1610094

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” 2018.

[8] M. Peng, Q. Zhang, Y.-g. Jiang, and X. Huang, “Cross-domain sentiment classification with target domain specific information,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 2505–2513. [Online]. Available: https://www.aclweb.org/anthology/P18-1233

[9] D. Bollegala, T. Mu, and J. Y. Goulermas, “Cross-domain sentiment classification using sentiment sensitive embeddings,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 2, pp. 398–410, Feb 2016.

[10] N. X. Bach, V. T. Hai, and T. M. Phuong, “Cross-domain sentiment classification with word embeddings and canonical correlation analysis,” in *Proceedings of the Seventh Symposium on Information and Communication Technology*, ser. SoICT ’16. New York, NY, USA: ACM, 2016, pp. 159–166. [Online]. Available: http://doi.acm.org/10.1145/3011077.3011104

[11] S. J. Pan, X. Ni, J.-T. Sun, Q. Yang, and Z. Chen, “Cross-domain sentiment classification via spectral feature alignment,” in *Proceedings of the 19th International Conference on World Wide Web*, ser. WWW ’10. New York, NY, USA: ACM, 2010, pp. 751–760. [Online]. Available: http://doi.acm.org/10.1145/1772690.1772767

[12] C.-K. Lin, Y.-Y. Lee, C.-H. Yu, and H.-H. Chen, “Exploring ensemble of models in taxonomy-based cross-domain sentiment classification,” in *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, ser. CIKM ’14. New York, NY, USA: ACM, 2014, pp. 1279–1288. [Online]. Available: http://doi.acm.org/10.1145/2661829.2662071

[13] D. Bollegala, D. Weir, and J. Carroll, “Cross-domain sentiment classification using a sentiment sensitive thesaurus,” *IEEE Transactions on*

TABLE V
PIVOTS

MODEL	Worst Model	Best Model
$G \rightarrow E$	jogo, não, jogar, bom	não recomendo, melhor, melhor jogo, gráficos
$E \rightarrow G$	prós, contras, conclusão, não, produto	excelente, prós contras, não, ótimo, ótima

TABLE VI
BERT RESULTS

MODEL	Accuracy	Precision	Recall	F-Measure
$E \rightarrow G$	0.750	0.760	0.750	0.745
$G \rightarrow E$	0.862	0.865	0.860	0.860

Knowledge and Data Engineering, vol. 25, no. 8, pp. 1719–1731, Aug 2013.

[14] X. Glorot, A. Bordes, and Y. Bengio, "Domain adaptation for large-scale sentiment classification: A deep learning approach," in *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ser. ICML'11. USA: Omnipress, 2011, pp. 513–520. [Online]. Available: http://dl.acm.org/citation.cfm?id=3104482.3104547

[15] S. Tan and X. Cheng, "Improving scl model for sentiment-transfer learning," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, ser. NAACL-Short '09. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009, pp. 181–184. [Online]. Available: http://dl.acm.org/citation.cfm?id=1620853.1620903

[16] E. A. C. Júnior, V. Q. Marinho, L. B. dos Santos, T. F. C. Bertaglia, M. V. Treviso, and H. B. Brum, "Pelesent: Cross-domain polarity classification using distant supervision," *CoRR*, vol. abs/1707.02657, 2017. [Online]. Available: http://arxiv.org/abs/1707.02657

[17] J. Blitzer, M. Dredze, and F. Pereira, "Biographies, bollywood, boomboxes and blenders: Domain adaptation for sentiment classification," in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Prague, Czech Republic: Association for Computational Linguistics, Jun. 2007, pp. 440–447. [Online]. Available: https://www.aclweb.org/anthology/P07-1056

[18] K. W. Church and P. Hanks, "Word association norms, mutual information, and lexicography," *Comput. Linguist.*, vol. 16, no. 1, pp. 22–29, Mar. 1990. [Online]. Available: http://dl.acm.org/citation.cfm?id=89086.89095

[19] Y. Ziser and R. Reichart, "Neural structural correspondence learning for domain adaptation," *CoRR*, vol. abs/1610.01588, 2016. [Online]. Available: http://arxiv.org/abs/1610.01588

[20] J. G. R. de Souza, A. de Paiva Oliveira, and A. Moreira, "Development of a brazilian portuguese hotel's reviews corpus," in *PROPOR*, 2018.