Biographies or Blenders: Which Resource Is Best for Cross-Domain Sentiment Analysis?

Conference Paper · March 2012

DOI: 10.1007/978-3-642-28604-9_40

2 authors:

Natalia Ponomareva
University of Wolverhampton

Mike Thelwall
University of Wolverhampton

Biographies or Blenders: Which Resource is Best for Cross-Domain Sentiment Analysis?

Natalia Ponomareva and Mike Thelwall

University of Wolverhampton, UK
Statistical Cybermetrics Research group,
{nata.ponomareva, m.thelwall}@wlv.ac.uk

Abstract. Domain adaptation is usually discussed from the point of view of new algorithms that minimise performance loss when applying a classifier trained on one domain to another. However, finding pertinent data similar to the test domain is equally important for achieving high accuracy in a cross-domain task. This study proposes an algorithm for automatic estimation of performance loss in the context of cross-domain sentiment classification. We present and validate several measures of domain similarity specially designed for the sentiment classification task. We also introduce a new characteristic, called domain complexity, as another independent factor influencing performance loss, and propose various functions for its approximation. Finally, a linear regression for modeling accuracy loss is built and tested in different evaluation settings. As a result, we are able to predict the accuracy loss with an average error of 1.5% and a maximum error of 3.4%.

性能下降的两个因素：
1、领域相似度；
2、领域复杂度

只是线性回归，还是说两者是线性的关系？

1 Introduction

Lack of annotated corpora that would suit the needs of NLP researchers is a common problem for many NLP tasks where machine learning is involved. More specifically, whilst there is a plethora of available annotated resources on the Internet, in many cases these resources do not match the data to be classified. Most of studies on domain adaptation research how to decrease the loss of performance when adapting a classifier trained on a domain with available annotated data to the target domain [6], [5], [7]. However, the choice of pertinent data seems to be equally important for obtaining satisfactory results. Indeed, machine-learning techniques are based on the assumption that training and test data are driven from the same probability distribution, and, therefore, they perform much better when training and test data sets are alike. Thus, the task of finding the best training data transforms into the task of finding data whose feature distribution is similar to the test one.

数据丰富度与领域不匹配

无法保证训练数据和测试数据是来自同一领域的（同分布的）

感觉是很多适应性工作的共同假设

The present paper considers this issue in the context of sentiment classification (SC). A drastic drop in performance when training and test data are different is common for cross-domain SC algorithms [12]. Usually this problem is tackled by proposing another domain-adaptation method, like ensembles of classifiers [2] or graph-based approaches [13]. However, these algorithms normally do not work well for very different source and target domains. Some studies also prove the benefit of using only the closest data set for training instead of exploiting all available data. For example, combinations of classifiers from different domains in some cases perform much worse than a single classifier trained on the closest domain [4]. These facts confirm the necessity to find a technique for choosing the most appropriate training data out of various available annotated data sets. The main goal of the present research is to analyse the principal factors causing performance loss and construct a model to predict the accuracy drop for a given pair of training and test data sets.

最好使用和测试数据接近的数据，而非所有数据

分析性能下降的因素，以及如何预测性能下降

Due to the specificity of SC, the most discriminative features used in machine learning are not necessary the most frequent words but words bearing sentiment. Numerous studies in Sentiment Analysis pointed out that adjectives, adverbs and verbs are usually good indicators of sentiment [10]. Thus, reduction of the feature set to unigrams and bigrams containing these parts-of-speech (POS) may give a better approximation to the real feature distribution. After building a feature representation of the domain we need to establish a similarity metric to measure closeness between source and target domain distributions. We analyse and compare different similarity functions, e.g. geometrically motivated measures and metrics borrowed from Information Theory and Corpus Linguistics (as the notion of domain similarity is identical to the notion of corpora comparability). Similarity in the sense of the proposed measures is controlled by the most frequent features, as they make the major impact on the value of the functions. At the same time, the tail of the distribution is also important because it indicates the complexity of the problem: longer tails correspond to richer domains which tend to be more complex for machine-learning tasks. We demonstrate this property in Section 3 by comparing in-domain accuracy for different corpora. Our experiments show that more diverse domains like books and DVDs give lower accuracy than kitchen appliances or electronics. In the light of this, we introduce another measure, called domain complexity, that is determined by the tail of the distribution and reflects the difficulty of classification task for a given data set. In Section 4.2 several functions for approximating domain complexity are suggested, and their high correlation with in-domain accuracy is evidenced.

减小特征集，只包含这些词性以及2元3元词

相似度由最频繁的特征决定

分布的尾部与复杂度

In the final step of our work we model cross-domain performance loss on the basis of two characteristics: domain similarity and complexity variance between source and target domains. We assume linear dependency between model output and its parameters and use multiple linear regression framework to compute model coefficients.

The paper is structured as follows. In Section 2 some related studies are overviewed. In Section 3 we describe our data and give preliminary results on domain adaptation. Section 4 introduces and validates measures of domain similarity and complexity, which are used in accuracy loss modeling described and evaluated in Sections 5-6. Finally, we sum up our contributions and give directions for the future research in Section 7.

2 Related work

The majority of works on cross-domain SC aim to elaborate a good transfer algorithm which minimises the performance loss observed on a target domain. The main approaches there include ensembles of classifiers [2], Structural Correspondence Learning [4], graph algorithms [13] and Spectral Feature Alignment [9]. However, there is very little research on evaluation of corpus quality for being a good training data for a given test data. In the paper of [4] the necessity of a domain similarity measure was mentioned for the first time. The authors proposed to apply $\mathcal{A}$-measure [3] and computed its proxy in a supervised manner. At the same time, a great interest to this problem has been expressed recently with regard to other NLP tasks, e.g. dependency parsing and POS tagging. During the last two years several works claiming the importance of domain similarity have been published [1], [11]. The number of recent parallel studies prove the timeliness of the present research.

The objective of the work described in [1] is similar to ours; the authors aimed to estimate the performance loss of a POS tagger across domains. They experimented with several POS tagger algorithms and different domains from BNC, provided by BNC annotators. Their conclusions are quite optimistic as they were able to predict the accuracy loss with 95% of confidence. However, we see several problems with the results reported in the paper. First of all, the best similarity measure was chosen on the basis of its correlation with the accuracy instead of the accuracy drop. In our opinion, this is not absolutely correct, because it implies the same accuracy for any pair of identical domains, which is not true as in-domain accuracy is determined by the corpus properties. Second, the authors scrutinised quite similar data, as the average cross-domain accuracy lies between 93-95% which is very close to the state-of-the-art accuracy for POS tagging. Therefore, their results can not be presumed to be representative and be generalised for the case of rather different domains.

批判了一番[1]，认为不正确、不可信。要用性能下降。

The study of [11] approached the cross-domain problem from a different angle – instead of choosing the best corpus out of the set of existing ones, the authors disregarded domain boundaries and proposed to acquire a corpus of separate documents similar to the test data. One of the interesting ideas suggested by this research is the use of topic models for document representation. The authors compared topic models representation with word representation and concluded the slight advantage of the former for the dependency parser problem, although the difference between them is significant only for one test data set out of the three considered in the paper.

从各个数据集中选数据。原来以前不这么做。

3 Preliminary results

Our data consist of Amazon product reviews on 7 different topics: books (BO), electronics (EL), kitchen&housewares (KI), DVDs (DV), music (MU), health&personal care (HE) and toys&games(TO) [4]. Reviews were rated using binary scale, 1-2 stars reviews are considered as negative and 4-5 stars reviews

- as positive. The data within each domain are balanced, they contain 1000 positive and 1000 negative reviews.

In Table 1 we can see that BO, DV and MU have the longest reviews, which also implies that the size of their dictionaries is larger than that of other domains. These observations confirm our intuition that vocabularies of BO and DV are more diverse and sophisticated.

Table 1. Reviews corpora statistics

corpus	num words	mean words	vocab size	vocab size ($freq \geq 3$)	% of rare words
BO	364k	181.8	23k	8,256	64.77
DV	397k	198.7	24k	8,632	64.16
MU	300k	150.1	19k	6,163	67.16
EL	236k	117.9	12k	4,465	61.71
KI	198k	98.9	11k	4,053	61.49
TO	206k	102.9	11k	4,018	63.37
HE	188k	93.9	11k	4,022	61.83

First, we accomplish in-domain experiments which represent the top boundary that any cross-domain algorithm is aiming to reach (Fig.1). For that a linear-kernel SVM classifier with a 5-fold cross-validation setup is used. We test different feature sets and conclude that unigrams together with bigrams of stems weighted with binary values yield the best performance for all domains under consideration. Fig.1 reveals that in-domain accuracies are higher for corpora with a smaller vocabulary and a lower percentage of rare words (words that appear less than 3 times in total). One of the logical explanations of this fact is that vocabulary diversity and long tail of feature distribution make automatic learning, in general, and SC, in particular, more difficult. 稀有词越多越难处理

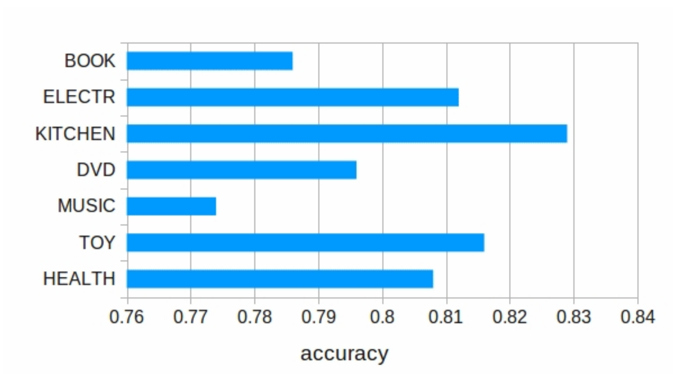

Fig. 1. In-domain accuracies

Fig.2 gives cross-domain accuracies for all domain pairs. Products on the y-axis are used as training and products on the x-axis - as test data. It can be seen that MU, DV and BO are relatively close and the same is true for HE, EL and KI. The adaptation of the classifier inside these domain clusters costs less than 5% in terms of accuracy drop for almost all domain pairs. At the same time, the drop of performance jumps sharply up to 14-15% for completely different domain pairs like (DV, HE) or (BO, EL). Interestingly, the performance loss is non-symmetric and it is normally higher when the source domain is more diverse. Non-symmetry of the drop is especially evident for a pair of very different domains.

性能下降不是对称的，两个领域分别作为测试集和训练集的结果并不相等。源领域约多样，在目标领域上下降越多

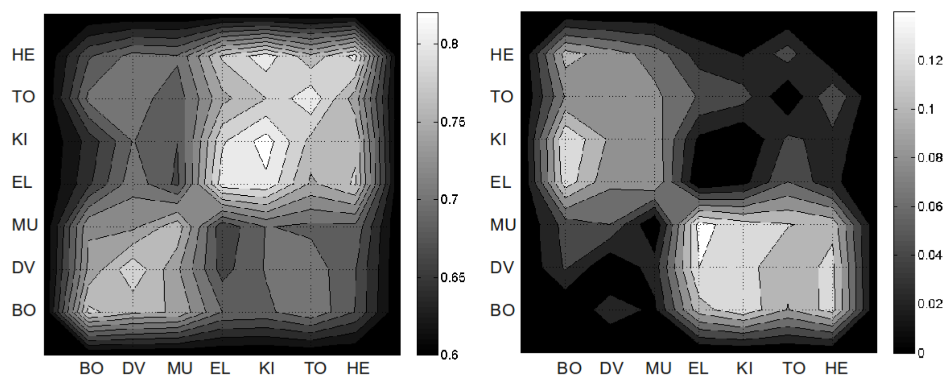

Fig. 2. Cross-domain results: a) accuracy, b) accuracy drop.

4 Domain similarity and complexity

4.1 Domain similarity

A domain is represented by a corpus of documents, therefore, domain similarity is identical to corpus similarity. The easiest way to find an appropriate measure of corpus similarity that would suit our needs is to adopt existing measures widely used in Corpus Linguistics (CL) for measuring the comparability of corpora pairs. However, there are certain differences between our objectives and CL ones:

语料库即领域

1. We are not interested in all terms of a corpus but rather on those which bear sentiment and can be considered as discriminative features for a machine learning algorithm. The study on SA suggested that adjectives, verbs and adverbs suit our purposes the best, therefore, we keep only unigrams and bigrams that contain those POS as features to compute corpus similarity.

2. CL compares frequencies of terms and, thus, the terms with the highest frequency bring the most substantial impact into similarity. However, it is not obvious that the most frequent terms are the most valuable for SC. Therefore,

用tf-idf而非词频

together with term frequencies we adopt and compare TF-IDF and IDF term weighting schemes.

不同的方法

Kilgarriff [8] introduced and tested 3 measures of corpus similarity - χ^2, Spearman rank correlation coefficient and cross-entropy, and in their experiments χ^2 showed the best correlation with the gold standard. We adopt χ^2 together with some measures borrowed from Information theory – Kullback-Leibler divergence (D_{KL}) and its symmetric analogue Jensen-Shannon divergence (D_{JS}), and other well-known similarity functions including the Jaccard coefficient (*Jaccard*) and cosine similarity (*cosine*). Note that some proposed functions measure actual similarity like cosine or Jaccard, while the others (χ^2, D_{KL} and D_{JS}) measure distance. This difference does not really matter much as similarity can always be obtained by inverting the distance.

Table 2 gives correlations between the accuracy drop and similarity functions applied to different feature representations of corpora. There "freq", "TFIDF" and "IDF" state for a term-weighting scheme and "filtr" means that the features were filtered by appearance of adjectives, adverbs and verbs. The different sign of the correlations is due to the fact that domain distance is directly-proportional and domain similarity is inversely-proportional to the accuracy loss.

Table 2. Correlation for different domain similarity measures

measure	R (freq)	R (filtr., freq)	R (filtr., TFIDF)	R (filtr., IDF)
cosine	-0.790	-0.840	-0.836	-0.863
Jaccard	-0.869	**-0.879**	**-0.879**	**-0.879**
χ^2	0.855	0.869	**0.876**	**0.879**
D_{KL}	0.734	0.827	0.676	0.796
D_{JS}	0.829	0.833	0.804	**0.876**

We can observe that correlation is higher for filtered features. A small increase in correlation can be identified for IDF weights with respect to frequencies and TFIDF. As far as similarity functions are concerned, strangely, the simplest function which does not depend on feature weights, Jaccard, shows the best overall correlation. χ^2 with IDF weights gives the same results and we select it for the accuracy loss modeling as it was previously chosen to be the most adequate measure of domain similarity.

Clusters of similar domains are clearly seen in Fig. 3 where pairwise similarities of our corpora according to $\chi^2_{inv} = \frac{1}{\chi^2}$ measure are depicted. Analysing this figure we can say that the boundary between similar and distinct domains approximately corresponds to $\chi^2_{inv} = 1.7$.

4.2 Domain complexity

Similarity between domains is mostly controlled by frequent words, but the shape of the corpus distribution is also influenced by rare words representing its tail. In Section 3 we showed that richer domains with more rare words are more complex

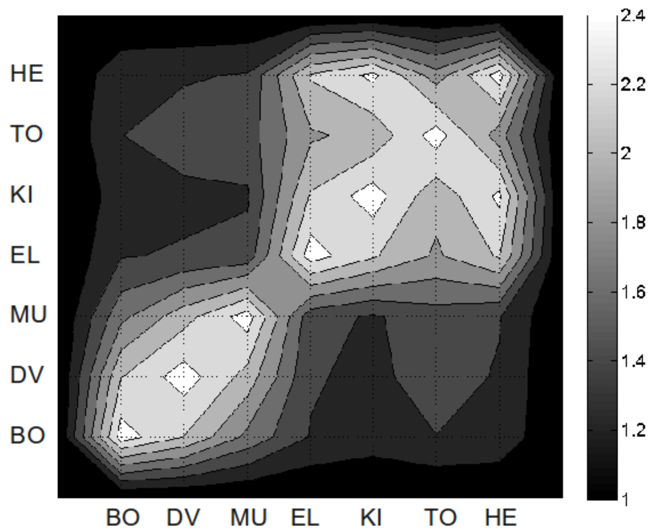

Fig. 3. Pairwise similarity of corpora according to χ^2_{inv}

for SC (Fig.1). We can also observe that the accuracy loss is higher in cross-domain settings when the source domain is more complex than the target one (Fig.2). These facts suggest that domain complexity in the sense of vocabulary diversity is another important characteristic regulating the performance loss.

这也就是为什么说领域复杂度和性能下降有关

We propose several measures to approximate domain complexity: *percentage of rare words*, *word richness*, calculated as a proportion of vocabulary size in a corpus size, and *relative entropy*, which is a percentage of corpus entropy out of the maximum entropy when all vocabulary words are distributed uniformly. Table 3 clearly divides all corpora into 2 groups: more complex domains like BO, DV and MU and simpler domains like EL, KI, TO and HE. More complex domains have higher number of rare words, a higher level of word richness and lower relative entropy. The last phenomena can be explained that distributions of more simple domains are closer to uniform than distributions of more complex domains.

如何评估复杂度

Table 4, where correlations between complexity measures and in-domain accuracy are presented, shows that the number of rare words approximates domain complexity the best. Interestingly, relative entropy, which is determined by the shape of the distribution and not only its tail, gives the lowest correlation with the in-domain accuracy. That means that domain complexity according to our definition is mostly associated with noisy low frequency terms.

哪种方法最好

Complexity is a property of one domain, therefore, for cross-domain settings we need to consider complexity variance between source and target domains. Let us denote Δc a measure of complexity variance which is positive for more complex and negative for more simple target domains:

$$\Delta c = c_t - c_s, \qquad (1)$$

where, c_s and c_t are relative entropy of source and target domains respectively.

Table 3. Corpora complexity

corpus	accuracy	% of rare words	word richness	rel.entropy, %
BO	0.786	64.77	0.064	9.23
DV	0.796	64.16	0.061	8.02
MU	0.774	67.16	0.063	8.98
EL	0.812	61.71	0.049	12.66
KI	0.829	61.49	0.053	14.44
TO	0.816	63.37	0.053	15.27
HE	0.808	61.83	0.056	15.82

Table 4. Pearson correlation with indomain accuracy for complexity measures

% of rare words	word richness	rel.entropy
-0.904	-0.846	0.793

5 Modeling the accuracy loss

In the previous section we presented two domain characteristics: domain similarity and complexity variance, that were proved to have an impact on the accuracy loss. These measures are independent, their values do not imply each other. In particular, if domains are different, they can still be of the same complexity level, or either of them can be more complex than another one. For example, on one hand, TO seems to be much more complex than EL, KI and HE (Table 3), but, on the other hand, it is quite similar to these domains (Fig.3).

复杂度和相似度无关

To model the performance drop we assume its linear dependency on domain similarity and complexity variance and propose the following linear regression model F:

$$F(s_{ij}, \Delta c_{ij}) = \beta_0 + \beta_1 s_{ij} + \beta_2 \Delta c_{ij}, \qquad (2)$$

where s_{ij} stands for domain similarity (or distance) between domains i and j and Δc_{ij} - for the difference between domain complexities. The unknown coefficients β_i are solutions of the following system of linear equations:

$$\beta_0 + \beta_1 s_{ij} + \beta_2 \Delta c_{ij} = \Delta a_{ij}, \qquad (3)$$

where Δa_{ij} is the accuracy drop when adapting the classifier from domain i to domain j.

Apart from establishing the unknown coefficients, it is necessary to accomplish a thorough analysis of model goodness. In particular, it is important to analyse whether the correlation between our predictors and the response variable is statistically significant and whether each of the predictors give a significant impact to the model. Another part of the evaluation consists of estimating the average accuracy and prediction of model behaviour on unseen examples.

6 Evaluation

Our data comprise 7 domains, which gives us a sample of 42 pairs to estimate unknown parameters and validate the model. We intended to artificially increase the number of examples by random division of each domain in 2 parts, but smaller data sets did not give reliable estimations for domain similarity and complexity. Therefore, we decided to carry out the experiments with the initial data. In previous section we established that χ^2 and the percentage of rare words are the best measures for domain similarity and complexity respectively. These functions were used in our accuracy loss modeling.

The evaluation of the constructed regression model includes following steps:

- *Global test (or F-test)* to verify statistical significance of regression model with respect to all its predictors.
- *Test on individual variables (or t-test)* to reveal regressors that do not bring a significant impact into the model.
- *Leave-one-out-cross validation* for the data set of 42 examples.
- *Domain-out validation* to estimate the accuracy drop obtained on domains which do not participate in modeling. This can give us a more reliable approximation to a real performance loss observed on unseen data.

6.1 Model validation

The null hypothesis for global test states that there is no correlation between regressors and the response variable. Our purpose is to demonstrate that this hypothesis must be rejected with a high level of confidence. In other words, we have to show that coefficient of determination R^2 is high enough to consider its value significantly different from zero.

Table 5. F-test for the linear regression model

R^2	R	F-value	p-value
0.873	0.935	134.60	«0.0001

Table 5 proves statistical significance of our model with the confidence level more than 99.9%, the correlation between the response and predictors is over

0.85. However, the global test evaluates the model in general, affirming that one of the coefficients is different from zero, but it does not say anything about each of the model parameters in particular.

Table 6. t-test for regression coefficients

	β_0	β_1	β_2
value	-8.67	27.71	-0.55
standard error	1.08	1.77	0.11
t-value	-8.00	15.67	-4.86
p-value	«0.0001	«0.0001	«0.0001

Table 6 presents the results of the test on individual coefficients. All of them are justified to be statistically significant with the confidence level higher than 99.9%. The model conforms with an intuitive assumption that high similarity leads to a small accuracy drop and vice versa. It also proves that relatively simpler source domains tend to give a smaller accuracy drop on relatively more complex target domains.

6.2 Accuracy estimation

The next step after validating the model is an estimation of the error it gives for unseen data. We evaluate the accuracy for 2 different settings: leave-one-out cross-validation and domain-out validation, which means that we take each domain out of the consideration when learning the model and use it only for testing. The latter procedure seems to give more reliable estimate for the accuracy on unseen data.

Results of leave-one-out cross-validation are presented in Table 7. The average error is around 1.5% and it does not exceed a threshold of 3.4% with the confidence level of 95%. Table 7 gives detailed results for different accuracy drops. As it can be seen, errors are very similar regardless of how close the domains are, but lower values are being noticed for more similar domains (with accuracy drops less than 5%). This is a strength of the model as our main purpose is to identify the closest domains.

Table 7. Leave-one-out cross-validation results

accuracy drop	standard error	standard deviation	max error, 95%
all data	1.566	1.091	3.404
$< 5\%$	1.465	1.133	3.373
$>5\%, < 10\%$	1.646	1.173	3.622
$> 10\%$	1.556	1.166	3.519

The results of domain-out validation reveal 3 domains (BO, MU and EL) with a very high standard error close to 2% (Table 8). It suggests that these domains have a higher level of noise and the size of the data is not large enough for reliable estimation of domain similarity and complexity. Concerning the rest of the domains, the constructed model gives much more accurate prediction of the performance loss with the standard error varying from 1.1% to 1.25%. This is much lower than the average error of leave-one-out cross-validation. The maximum error for the confidence level equal to 95% is high for all domains because of the small data sample containing only 12 examples. To obtain a more precise confidence interval for the error rate, more domain pairs are needed.

Table 8. Domain-out validation results

Domain	correlation coeff.	standard error	standard deviation	max error, 95%
BO	0.882	2.0856	1.400	4.600
DV	0.943	1.253	1.308	3.602
MU	0.898	1.789	1.869	5.145
EL	0.946	1.861	1.156	3.937
KI	0.970	1.216	0.750	2.563
TO	0.932	1.148	1.381	3.628
HE	0.974	1.111	1.160	3.195

7 Conclusions and future work

The paper gives an alternative insights into the domain adaptation problem. Instead of proposing a new efficient technique of domain transfer we present the method of modeling the performance loss on different data sets. This technique can help to choose the most appropriate data similar to the existing test one. First, we introduce measures of domain similarity and complexity and demonstrate their influence into the performance loss of a cross-domain classifier. Then, using these measures, we construct a linear regression model, which we verify and test in different evaluation settings. Our model demonstrates a satisfactory behaviour, predicting performance loss with an error of 1.5%. The standard error and deviation of the predictions are slightly lower for small accuracy drops that proves the ability of the model to identify similar domains.

In future, we plan to carry out experiments on larger data sets, verifying our results on new topics and genres. In particular, challenging data gathered from Twitter, MySpace and Youtube will be examined and the possibility to use them in the cross-domain task will be studied. We believe that these experiments will prove the importance of the domain complexity measure because corpus complexity differs even more when considering distinct genres than just different topics.

Another direction of the future work will focus on improvement of proposed measures of domain similarity and complexity. The work of [11] shows that for some data sets topic models are more efficient for approximation of domain similarity than word representations. However, the authors dealt with dependency parsing and their conclusions may not be valid for SC. Therefore, we intend to compare topic models with our unigrams+bigrams corpus representations and analyse their influence into the overall accuracy of our prediction model.

Acknowledgements

This work was supported by a European Union grant by the 7th Framework Programme, Theme 3: Science of complex systems for socially intelligent ICT. It is part of the CyberEmotions project (contract 231323).

References

1. Asch, V.V., Daelemans, W.: Using domain similarity for performance estimation. In: Proceedings of the 2010 Workshop on Domain Adaptation for NLP, ACL'10. pp. 31–36 (2010)
2. Aue, A., Gamon, M.: Customizing sentiment classifiers to new domains: A case study. In: Proceedings of RANLP'05 (2005)
3. Ben-David, S., Blitzer, J., Crammer, K., Pereira, F.: Analysis of representations for domain adaptation. In: Advances in Neural Information Processing Systems (NIPS) (2006)
4. Blitzer, J., Dredze, M., Pereira, F.: Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In: Proceedings of ACL'07. pp. 440–447 (2007)
5. Blitzer, J., McDonald, R., Pereira, F.: Domain adaptation with structural correspondence learning. In: Proceedings of EMNLP'06. pp. 120–128 (2006)
6. Daume III, H., Marcu, D.: Domain adaptation for statistical classifiers. Artificial Intelligence Research 26, 101–126 (2006)
7. Glorot, X., Bordes, A., Bengio, Y.: Domain adaptation for large-scale sentiment classification: A deep learning approach. In: Proceedings of ICML'11 (2011)
8. Kilgarriff, A.: Comparing corpora. International Journal of Corpus Linguistics 6(1), 97–133 (2001)
9. Pan, S.J., Niz, X., Sunz, J.T., Yangy, Q., Chen, Z.: Cross-domain sentiment classification via spectral feature alignment. In: Proceedings of WWW2010 (2010)
10. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval 2(1-2), 1–135 (2008)
11. Plank, B., van Noord, G.: Effective measures of domain similarity for parsing. In: Proceedings of ACL'11. pp. 1566–1576 (2011)
12. Read, J.: Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In: Proceedings of the ACL Student Research Workshop. pp. 43–48 (2005)
13. Wu, Q., Tan, S., Cheng, X.: Graph ranking for sentiment transfer. In: Proceedings of ACL-IJCNLP'09. pp. 317–320 (2009)