Department: Affective Computing and Sentiment Analysis
Editor: Erik Cambria, Nanyang Technological University

Cross-Lingual Sentiment Quantification

Andrea Esuli

Alejandro Moreo

Fabrizio Sebastiani
Consiglio Nazionale delle Ricerche, Italy

Abstract—Sentiment Quantification **is the task of estimating the relative frequency of sentiment-related classes — such as Positive and Negative — in a set of unlabeled documents. It is an important topic in sentiment analysis, as the study of sentiment-related quantities and trends across a population is often of higher interest than the analysis of individual instances. In this work, we propose a method for *Cross-Lingual Sentiment Quantification*, the task of performing sentiment quantification when training documents are available for a source language $\mathcal{S}$ but not for the target language $\mathcal{T}$ for which sentiment quantification needs to be performed. Cross-lingual sentiment quantification (and cross-lingual *text* quantification in general) has never been discussed before in the literature; we establish baseline results for the binary case by combining state-of-the-art quantification methods with methods capable of generating cross-lingual vectorial representations of the source and target documents involved. Experiments on publicly available datasets for cross-lingual sentiment classification show that the presented method performs cross-lingual sentiment quantification with high accuracy.**

In Cross-Lingual Text Classification, documents may be expressed in either a *source* language $\mathcal{S}$ or a *target* language $\mathcal{T}$, and training documents are available only for $\mathcal{S}$ but not for $\mathcal{T}$; cross-lingual text classification thus consists of leveraging the training documents in the source language in order to train a classifier for the target language, also using the fact that the classification scheme $\mathcal{C}$ is the same for both $\mathcal{S}$ and $\mathcal{T}$. Cross-lingual text classification has been widely investigated in the literature [1], [2].

A companion task which instead has never been tackled, and which is the object of this paper, is *Cross-Lingual Text Quantification*, the task of performing “quantification” across a source language $\mathcal{S}$ and a target language $\mathcal{T}$. *Quantification* is a supervised learning task that consists of predicting, given a set of classes $\mathcal{C}$ and a set D (a *sample*) of unlabeled items drawn from some domain $\mathcal{D}$, the *prevalence* (i.e., relative frequency) $p_c(D)$ of each class $c \in \mathcal{C}$ in D.

Put it another way, given an unknown distribution $p_{\mathcal{C}}(D)$ of the members of D across $\mathcal{C}$ (the *true distribution*), quantification consists in generating a *predicted distribution* $\hat{p}_{\mathcal{C}}(D)$ that approximates $p_{\mathcal{C}}(D)$ as accurately as possible [3].

Quantification is especially important for application fields characterized by an interest in aggregate (rather than individual) data, such as the social sciences, market research, political science, and epidemiology. These disciplines often face the need to label data in highly dynamic scenarios [4], i.e., scenarios in which the distribution of data in the unlabeled set may be very different from the distribution of data in the training set. In such contexts, accurate class prevalence estimation may be challenging, due to the fact that the "iid assumption" on which standard learning methods are based (i.e., the assumption that the training set and the test set are identically and independently sampled from the *same* data distribution) is obviously not verified.

Sentiment quantification [5] is the task of interest in all contexts in which the results of sentiment analysis are to be analyzed at the aggregate level. For instance, hardly anyone among those who perform sentiment analysis for Twitter data are interested in determining the sentiment conveyed by a single tweet; in most such applications, figuring out *the percentage* and *the intensity* [6] of tweets that exhibit a certain sentiment is the real goal, which shows that quantification (and not classification) should be the task to focus on [7]. This paper adds cross-linguality to the picture, thus addressing those application contexts characterized by the absence of training data for the "target" language of interest, and the presence of training data for a different "source" language [8]. Everything we say in this paper straightforwardly extends to dealing with the simultaneous presence of several source languages and/or several target languages.

In principle, quantification can be straightforwardly solved via classification, i.e., by training a classifier h using training data labeled according to $\mathcal{C}$, classifying the unlabeled data in D via h, and counting, for each $c \in \mathcal{C}$, how many items in D have been attributed to c (the "classify and count" method).

However, research has conclusively shown [9], [10], [11], [12] that this approach leads to suboptimal quantification accuracy. To see this consider that a binary classifier h_1 for which $\text{FP} = 20$ and $\text{FN} = 20$ (FP and FN standing for the "false positives" and "false negatives", respectively, that it has generated on a given dataset) is worse, in terms of classification accuracy, than a classifier h_2 for which, on the same dataset, $\text{FP} = 18$ and $\text{FN} = 20$. However, h_1 is intuitively a better binary quantifier than h_2; indeed, h_1 is a perfect quantifier, since FP and FN are equal and thus, when it comes to class frequency estimation, compensate each other, so that the distribution of the test items across the class and its complement is estimated perfectly. Since classification and quantification pursue different goals, quantification should be tackled as a task of its own, using different evaluation measures and, as a result, different learning algorithms.

In this paper, we establish baseline results for (binary) cross-lingual sentiment quantification by combining a number of quantification methods with state-of-the-art cross-lingual projection methods.For performing this latter task we explore *Structural Correspondence Learning* (SCL [1]) and *Distributional Correspondence Indexing* (DCI [2]), since (i) SCL is arguably the most representative cross-lingual projection method in the literature (and thus a mandatory baseline in lab experiments of related research), while DCI is a cross-lingual projection method that has recently demonstrated state-of-the-art performance in cross-lingual text classification [13], and (ii) both methods provide a general procedure for projecting source and target documents onto a common vector space, and (iii) the code implementing both methods is publicly available and easily modifiable. Other cross-lingual methods proposed in the literature learn representations that are dependent on the set of unlabeled documents to classify (in lab experiments: the test set). This implicitly means that each new unlabeled set to quantify upon would require retraining from scratch, something that would prove prohibitive in the experimental setting of quantification.

METHOD

Different quantification methods have been proposed that exploit the classification outcomes that a previously trained classifier delivers on unlabeled data. We explore different cross-lingual sentiment quantification methods that result from the combination of a cross-lingual projection method, a "classify and count" policy, and an estimate correction method. In this paper, we only address the binary case, where the classes {Positive,Negative} are indicated as $\mathcal{C} = \{\oplus, \ominus\}$.

Cross-Lingual Document Representations

In cross-lingual applications, SCL and DCI rely on the concept of *pivot term* (or simply *pivot*) [14] in order to bridge the gap between the different feature spaces which the different languages generate. In such contexts, pivots are defined as highly predictive pairs of translation-equivalent terms which behave in a similar way in their respective languages. Typical examples of pivots for sentiment-related applications are adjectives with domain-independent meaning such as "excellent" or "poor", and partially domain-dependent terms such as "fancy" (as found, e.g., in the arts and crafts domain and in the clothing domain) or "masterpiece" (as found, e.g., in the book domain, movie domain, and music domain), with their respective translations in other languages.

A common strategy to select the pivots automatically consists of taking the top elements from a list of terms ranked according to their mutual information to the label representing the domain (as computed from source-language training data), and filtering out those candidates whose translation equivalent shows a substantial prevalence drift in the target language. A word translation oracle, with a fixed budget of allowed calls, is assumed available.

Once pivots are selected, different methods can be defined in order to produce cross-lingual vectorial representations. Both SCL and DCI first represent documents as vectors $\mathbf{x}$ in a (weighted) bag-of-words model of dimension $|V|$ (with V being the vocabulary), and then apply a linear projection (parameterized by a matrix $\theta \in \mathbb{R}^{|V|L}$) of type $\mathbf{x}^\top\theta$, thus mapping $|V|$-dimensional vectors into L-dimensional vectors in a cross-lingual latent space.

To achieve this, the unlabeled collections from the source and target domains are inspected. The matrix can be subsequently used to project source documents (to train a classifier) and target documents (to classify them).

SCL builds the projection matrix by resolving an auxiliary prediction problem for each pair of translation-equivalent pivot terms. Each problem consists of predicting the presence of a pivot term based on the observation of the other terms. By solving the auxiliary problems (via linear classification), structural correspondences among terms and pivots are captured and collected as a matrix of correlations. This matrix is later decomposed using truncated SVD to generate the final projection matrix θ. DCI relies instead on the distributional hypothesis to directly model correspondences between terms and pivots. Each row of the projection matrix DCI computes represents a term profile, where each dimension quantifies the degree of correspondence (as measured by a *distributional correspondence function*) of the term to a pivot.

Classifying and Counting

An obvious way to solve quantification is by aggregating the scores assigned by a classifier to the unlabeled documents.

In connection to each of SCL and DCI we experiment with two different aggregation methods, one that uses a "hard" classifier (i.e., a classifier $h_\oplus : \mathcal{D} \rightarrow \{0, 1\}$ that outputs binary decisions, 0 for $\ominus$ and 1 for $\oplus$) and one that uses a "soft" classifier (i.e., a classifier $s_\oplus : \mathcal{D} \rightarrow [0, 1]$ that outputs posterior probabilities $\Pr(\oplus|\mathbf{x})$, representing the probability that the classifier attributes to the fact that $\mathbf{x}$ belongs to the $\oplus$ class). Of course, $\Pr(\ominus|\mathbf{x}) = (1 - \Pr(\oplus|\mathbf{x}))$.

The (trivial) *classify and count* (CC) quantifier then comes down to computing

$$\hat{p}^{\text{CC}}_\oplus(D) = \frac{\sum_{\mathbf{x}\in D} h_\oplus(\mathbf{x})}{|D|} \tag{1}$$

while the *probabilistic classify and count* quantifier (PCC [10]) is defined by

$$\hat{p}^{\text{PCC}}_\oplus(D) = \frac{\sum_{\mathbf{x}\in D} s_\oplus(\mathbf{x})}{|D|} \tag{2}$$

Of course, for any method M we have $\hat{p}^M_\ominus(D) = (1 - \hat{p}^M_\oplus(D))$.

Adjusting the Results of Classify and Count

A popular quantification method consists of applying an *adjustment* to the prevalence $\hat{p}_{\oplus}(D)$ estimated via "classify and count". It is easy to check that, in the binary case, the true prevalence $p_{\oplus}(D)$ and the estimated prevalence $\hat{p}_{\oplus}(D)$ are such that

$$p_{\oplus}(D) = \frac{\hat{p}^{\mathrm{CC}}_{\oplus}(D) - fpr_h}{tpr_h - fpr_h} \tag{3}$$

where tpr_h and fpr_h stand for the *true positive rate* and *false positive rate* of the classifier $h_{\oplus}$ used to obtain $\hat{p}^{\mathrm{CC}}_{\oplus}$. The values of tpr_h and fpr_h are unknown, but can be estimated via k-fold cross-validation on the training data. In the binary case this comes down to using the results $h_{\oplus}(\mathbf{x})$ obtained in the k-fold cross-validation (i.e., $\mathbf{x}$ ranges on the training documents) in equations

$$\hat{tpr}_h = \frac{\sum_{\mathbf{x}\in\oplus} h_{\oplus}(\mathbf{x})}{|\{\mathbf{x}\in\oplus\}|} \qquad \hat{fpr}_h = \frac{\sum_{\mathbf{x}\in\ominus} h_{\oplus}(\mathbf{x})}{|\{\mathbf{x}\in\ominus\}|} \tag{4}$$

We obtain estimates of $p^{\mathrm{ACC}}_{\oplus}(D)$, which define the *adjusted classify and count* method [12] (ACC) by replacing tpr_h and fpr_h in Equation 3 with the estimates of Equation 4, i.e.,

$$\hat{p}^{\mathrm{ACC}}_{\oplus}(D) = \frac{\hat{p}^{\mathrm{CC}}_{\oplus}(D) - \hat{fpr}_h}{\hat{tpr}_h - \hat{fpr}_h} \tag{5}$$

If the soft classifier $s_{\oplus}(\mathbf{x})$ is used in place of $h_{\oplus}(\mathbf{x})$, analogues of $\hat{tpr}_h$ and $\hat{fpr}_h$ from Equation 4 can be defined as

$$\hat{tpr}_s = \frac{\sum_{\mathbf{x}\in\oplus} s_{\oplus}(\mathbf{x})}{|\{\mathbf{x}\in\oplus\}|} \qquad \hat{fpr}_s = \frac{\sum_{\mathbf{x}\in\ominus} s_{\oplus}(\mathbf{x})}{|\{\mathbf{x}\in\ominus\}|} \tag{6}$$

We obtain $p^{\mathrm{PACC}}_{\oplus}(D)$ estimates, which define the *probabilistic adjusted classify and count* method (PACC [10]), by replacing all factors in the right-hand side of Equation 5 with their "soft" counterparts from Equations 2 and 6, i.e.,

$$\hat{p}^{\mathrm{PACC}}_{\oplus}(D) = \frac{\hat{p}^{\mathrm{PCC}}_{\oplus}(D) - \hat{fpr}_s}{\hat{tpr}_s - \hat{fpr}_s} \tag{7}$$

ACC and PACC define two simple linear adjustments to the aggregated scores of general-purpose classifiers.

We also investigate the use of a more recently proposed adjustment method beased on deep learning, called QuaNet [11]. QuaNet models a neural *non-linear* adjustment by taking as input all estimated prevalences from Equations 1, 2, 5, 7 (i.e., $\hat{p}^{\mathrm{CC}}_{\oplus}$, $\hat{p}^{\mathrm{ACC}}_{\oplus}$, $\hat{p}^{\mathrm{PCC}}_{\oplus}$, $\hat{p}^{\mathrm{PACC}}_{\oplus}$), several statistics (the $\hat{tpr}_h$, $\hat{fpr}_h$, $\hat{tpr}_s$, $\hat{fpr}_s$ estimates from Equations 4 and 6), the posterior probabilities $\Pr(\oplus|\mathbf{x})$ for each document $\mathbf{x}$, and the document vectors themselves. QuaNet relies on a recurrent neural network to produce "quantification embeddings" (i.e., dense, multi-dimensional representations of the information relevant to quantification observed from the input data), which are then used to generate the final prevalence estimates.

EXPERIMENTS

We tested each of the $2 \times 5 = 10$ combinations resulting from 2 approaches to generating cross-lingual projections (SCL and DCI) and 5 approaches to performing quantification (CC, PCC, ACC, PACC, and Quanet). The code to replicate all these experiments is available from GitHub[1]. Note that a dataset for sentiment classification is also a dataset for sentiment quantification, since one can compute the true class prevalences $p_{\oplus}(D)$ and $p_{\ominus}(D)$ by simply counting the assigned labels.

System setup

We use the NUT package[2] for SCL and the PYDCI[3] package [13] for DCI in order to generate the vectorial representations of all training and test documents. As the hard classifiers, we stick to the ones used by the original proponents of SCL and DCI, i.e., a linear classifier trained via Elastic Net [15] (implemented via the BOLT package[4]) for SCL, and a linear classifier trained via SVMs (implemented via the SCIKIT-LEARN package [16]) for DCI. As the soft classifier we instead use one trained via logistic regression (in its SCIKIT-LEARN implementation) for both SCL and DCI, since such classifiers are known to return "well-calibrated" posterior probabilities.

[1] http://github.com/HLT-ISTI/cl-quant
[2] http://github.com/pprett/nut
[3] http://github.com/HLT-ISTI/pydci
[4] http://github.com/pprett/bolt

The last point is fundamental for Equations 2, 6, 7 to return accurate values, since "well calibrated probabilities" is essentially a synonym of "good-quality probabilities". Posterior probabilities $\Pr(c|\mathbf{x})$ are said to be *well calibrated* when, given a sample D drawn from some population,

$$\lim_{|D|\rightarrow\infty}\frac{|\{\mathbf{x}\in c|\Pr(c|\mathbf{x})=\alpha\}|}{|\{\mathbf{x}\in D|\Pr(c|\mathbf{x})=\alpha\}|}=\alpha.$$

Intuitively, this property implies that, as the size of the sample D goes to infinity, e.g., 90% of the documents $\mathbf{x}\in D$ that are assigned a well calibrated posterior probability $\Pr(c|\mathbf{x}) = 0.9$ belong to class c. Some classifiers (e.g., those trained via logistic regression [17]) are known to return well calibrated probabilities. The posterior probabilities returned by some other classifiers (e.g., those trained via naïve Bayesian methods [18]) are known instead to be not well calibrated. Yet some other classifiers (e.g., those trained via SVMs) do not return posterior probabilities, but generic confidence scores. In these two last cases it is possible to map the obtained posterior probabilities / confidence scores into well calibrated posterior probabilities by means of some "calibration" method [19], [17].

We set all the hyper-parameters in SCL (number m of pivots, minimum support frequency ϕ for pivot candidates, dimensionality k of the cross-lingual representation, and the Elastic Net coefficient α) to ($m = 450$, $\phi = 30$, $k = 100$, $\alpha = 0.85$), i.e., to the values found optimal in previous literature [1] when optimizing for the German book review task. Along with previous work [13], in DCI we set the number of pivots and minimum support to $m = 450$ and $\phi = 30$. The dimensionality is $k = 450$ by definition, since in DCI each pivot corresponds to a dimension. In preliminary experiments we had used the same value $k = 450$ both for DCI and SCL, on grounds of "fairness". The results for SCL were slightly worse with respect to using $k = 100$; for SCL we thus decided to stick to the $k = 100$ value originally used by the creators of SCL [1]. As the distributional correspondence function we use cosine, since it is the best performer in previously published experiments [13]. For each setup we independently optimize the parameter C (which controls the regularization strength in the SVM and in the logistic regressor) via grid search in the log space defined by $C \in \{10^i\}_{i=-5}^{5}$, and via 5-fold cross-validation. The classifiers with the optimized hyper-parameters are then used in a 10-fold cross-validation run on the training data to produce the $\hat{tpr}_h$ and $\hat{fpr}_h$ estimates.

For the neural correction of QuaNet we use its publicly available implementation linked from the original paper[5]. We optimize the hyper-parameters of QuaNet using the German book review task (as done by Prettenhofer and Stein [1]); we end up using 64 hidden units in the recurrent cell of a two-layer stacked bidirectional LSTM, 1024 and 512 hidden units in the next-to-last feed-forward layers, and a drop probability of 0. We set the rest of the parameters to the same values as in the original QuaNet paper [11].

Experimental setting

We use the Webis-CLS-10 dataset [1] as the benchmark for our experiments. Webis-CLS-10 is a dataset originally proposed for cross-lingual sentiment classification experiments, and consisting of Amazon product reviews written in four languages (English, German, French, Japanese) and concerning three product domains (Books, DVDs, Music). There are 2,000 training documents, 2,000 test documents, and a number of unlabeled documents ranging from 9,000 to 50,000 for each combination of language and domain. The examples of $\oplus$ and $\ominus$ (which indicate positive and negative sentiment, resp.) are perfectly balanced (i.e., 50% each) in all sets (training, test, unlabeled). Following a consolidated practice in cross-lingual text classification, we always use English as the source language. We use the preprocessed version of the dataset[6], where terms correspond to uni-grams.

As the measures of quantification error we use *Absolute Error* (AE), *Relative Absolute Error* (RAE), and the *Kullback-Leibler Divergence*

[5] http://github.com/HLT-ISTI/quanet

[6] http://uni-weimar.de/medien/webis/corpora/corpus-webis-cls-10/cls-acl10-processed.tar.gz

(KLD), defined as:

$$\mathrm{AE}(p, \hat{p}, D) = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} |\hat{p}_c(D) - p_c(D)| \quad (8)$$

$$\mathrm{RAE}(p, \hat{p}, D) = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \frac{|\hat{p}_c(D) - p_c(D)|}{p_c(D)} \quad (9)$$

$$\mathrm{KLD}(p, \hat{p}, D) = \sum_{c \in \mathcal{C}} p_c(D) \log \frac{p_c(D)}{\hat{p}_c(D)} \quad (10)$$

since they are the most frequently used measures for evaluating quantification error [20].

The evaluation of a quantifier cannot be carried out on the basis on one single set of test documents. The reason is that, while in text classification experiments a test set consisting of n documents enables the evaluation of n different decision outcomes, in quantification the same test set would only allow to validate one single prevalence prediction. In order to allow statistically significant comparisons, Forman [12] proposed to run quantification experiments on a set of test samples, randomly sampled from the original set of test documents at different prevalence levels. Along with Forman [12], as the range of prevalences for the $\oplus$ class we use $\{0.01, 0.05, 0.10, \ldots, 0.90, 0.95, 0.99\}$. Similarly to previous work [11], we generate 100 random samples for each of the 21 prevalence levels, and report quantification error as the average across $21 \times 100 = 2100$ test samples. All samples consist of 200 documents. For each target language (German, French, Japanese) and product domain (Books, DVD, Music) the samples are the same across the different methods, which will enable us to evaluate the statistical significance of the differences in performance; to this aim, we rely on the non-parametric Wilcoxon signed-rank test on paired samples.

For each combination of target language and product domain, Table 1 reports quantification error (for each CLTQ method and for each evaluation measure) as an average across the 2100 test samples; we recall that English is always used as the source language, so that, e.g., the "German Books" experiment is about training on English book reviews and testing on German book reviews. Since QuaNet depends on a stochastic optimization, Table 1 reports the average and standard deviation across 10 runs.

Results

Overall, the results indicate that the combination DCI+PACC is the best performer in terms of AE and RAE, while DCI+QuaNet seems to behave slightly better in terms of KLD. Given recent theoretical results on the properties of evaluation measures for quantification [20], that indicate that AE and RAE are to be preferred to KLD, this leads us to prefer DCI+PACC.

A substantial superiority of DCI over SCL, as witnessed by the fact that, for each combination of evaluation measure, target language, and domain, the best performer always uses DCI and not SCL. This confirms previous results [2] that showed the superiority of DCI over SCL in monolingual sentiment classification contexts.

In both SCL and DCI the "hard" classifier tends to work comparatively better than the "soft" logistic regressor, as indicated by the fact that CC tends to outperform PCC and ACC tends (with some exceptions) to outperform PACC. As expected, ACC (the "adjusted" version of CC) performs substantially better than CC in all cases. What comes as a surprise, though, is the fact that the remarkable benefit PACC brings about in DCI with respect to its unadjusted variant PCC, is not consistently mirrored in the case of SCL (where the effect of adjusting is instead harmful, and especially so in terms of KLD).

The neural, non-linear adjustment of QuaNet, when applied to DCI vectors, performs somehow similarly to the best performer in several cases, and actually delivers the lowest average KLD error. That QuaNet does not perform as well with SCL can be explained by two facts (which are not independent of each other), i.e., the importance of the estimated posterior probabilities within QuaNet, and the suboptimal ability (as shown by the PCC and PACC results) in delivering accurate posterior probabilities for SCL vectors that the logistic regressor has shown.

CONCLUSION

The experiments we have performed show that structural correspondence learning (SCL) and distributional correspondence indexing (DCI), two previously proposed methods for cross-lingual text classification, can effectively be used in cross-lingual text quantification, a task that had never been tackled before in the literature.

Table 1. Cross-lingual sentiment quantification results for Webis-CLS-10. Boldface indicates the best result. Superscripts † and †† denote the method (if any) whose score is not statistically significantly different from the best one at $\alpha = 0.05$ (†) or at $\alpha = 0.005$ (††).

	Target		SCL					DCI				
	Language	Domain	CC	ACC	PCC	PACC	QuaNet	CC	ACC	PCC	PACC	QuaNet
AE	German	Books	0.092	0.040	0.237	0.375	0.203 (±0.006)	0.090	0.037	0.119	**0.027**	0.030 (±0.002)
	German	DVDs	0.104	0.045	0.221	0.331	0.178 (±0.009)	0.086	0.030	0.147	**0.028**	0.030 (±0.003)††
	German	Music	0.097	0.037††	0.151	0.101	0.072 (±0.007)	0.078	0.037††	0.109	0.039††	**0.030** (±0.002)
	French	Books	0.098	0.037	0.202	0.288	0.151 (±0.007)	0.098	0.038	0.122	**0.025**	0.036 (±0.003)
	French	DVDs	0.110	0.056	0.174	0.113	0.072 (±0.002)	0.091	0.037	0.117	**0.027**	0.045 (±0.005)
	French	Music	0.119	0.060	0.178	0.090	0.072 (±0.001)	0.074	0.030	0.160	**0.024**	0.047 (±0.010)
	Japanese	Books	0.127	0.072	0.194	0.124	0.095 (±0.002)	0.117	**0.060**	0.174	0.064	0.073 (±0.003)
	Japanese	DVDs	0.131	0.079	0.329	0.485	0.270 (±0.005)	0.104	0.045	0.128	**0.037**	0.058 (±0.006)
	Japanese	Music	0.118	0.059	0.242	0.377	0.228 (±0.007)	0.092	0.029	0.161	**0.027**	0.044 (±0.009)
	Average		0.111	0.054	0.214	0.254	0.149	0.092	0.038	0.138	**0.033**	0.044
RAE	German	Books	0.888	0.164	0.878	0.807	0.513 (±0.015)	1.135	0.246	1.411	**0.136**	0.248 (±0.034)
	German	DVDs	1.086	0.267	1.047	0.733	0.428 (±0.031)	1.070	0.223	1.709	**0.144**	0.234 (±0.020)††
	German	Music	1.056	0.194†	1.364	0.268	0.216 (±0.011)	0.947	0.194††	1.310	**0.153**	0.245 (±0.022)††
	French	Books	1.021	0.313	1.041	0.666	0.383 (±0.025)	1.227	0.407	1.426	**0.159**	0.330 (±0.026)
	French	DVDs	1.307	0.682	1.642	0.475	0.543 (±0.019)	0.938	0.176	1.284	**0.144**	0.223 (±0.016)
	French	Music	1.310	0.496	2.099	1.181	0.817 (±0.026)	0.834	**0.138**	1.803	0.208	0.276 (±0.039)†
	Japanese	Books	1.423	0.781	2.287	1.572	1.122 (±0.026)	1.196	**0.450**	1.935	0.639	0.570 (±0.032)
	Japanese	DVDs	1.392	0.785	0.833	0.947	0.557 (±0.012)	1.097	0.292	1.380	**0.213**	0.350 (±0.021)
	Japanese	Music	1.232	0.304	0.910	0.806	0.527 (±0.016)	0.973	**0.175**	1.800	0.198†	0.293 (±0.034)
	Average		1.191	0.443	1.345	0.828	0.567	1.046	0.256	1.562	**0.222**	0.308
KLD	German	Books	0.041	0.016	0.194	1.778	0.274 (±0.043)	0.040	0.032	0.062	0.028	**0.007** (±0.001)
	German	DVDs	0.050	0.013	0.172	0.987	0.139 (±0.034)	0.038	0.019	0.086	0.028	**0.007** (±0.001)
	German	Music	0.045	0.017††	0.090	0.062	0.027 (±0.005)	0.032	0.046	0.054	0.072	**0.008** (±0.001)
	French	Books	0.046	0.010††	0.146	0.748	0.115 (±0.024)	0.046	0.014	0.064	0.014	**0.010** (±0.001)
	French	DVDs	0.055	0.019	0.111	0.055	0.029 (±0.001)	0.040	0.012	0.060	**0.008**	0.012 (±0.002)
	French	Music	0.062	0.021	0.114	0.040	0.028 (±0.000)	0.030	0.040	0.097	**0.007**	0.014 (±0.004)
	Japanese	Books	0.068	0.028	0.132	0.065	0.043 (±0.001)	0.060	**0.020**	0.110	0.024	0.029 (±0.002)
	Japanese	DVDs	0.071	0.033	0.376	5.133	0.250 (±0.013)	0.051	0.014	0.069	**0.011**	0.020 (±0.003)
	Japanese	Music	0.061	0.022	0.202	1.629	0.234 (±0.024)	0.042	0.011	0.098	**0.009**	0.013 (±0.004)
	Average		0.055	0.020	0.171	1.166	0.127	0.042	0.023	0.078	0.022	**0.013**

The tested methods yield quantification predictions that are fairly close to the true prevalence; in terms of absolute error (arguably the most easy-to-interpret error criterion), and on average, the class prevalences predicted by DCI+PACC differ from the true prevalences by a margin of 3.3% on average, while this difference is 5.4% for SCL+ACC.

These results are encouraging, especially if we consider the fact that the quantifier is trained on a language different from the one on which quantification is performed (for which no training data are assumed to exist), and that a range of true prevalences different (and even extremely different) from the ones of the training set are tested upon.

Note also that these results are a further confirmation of the fact that, when our interest in automatically labeled data is at the aggregate level only (and not at the individual level), using "real" quantification methods (instead of standard classification methods in a "classify and count" fashion) is the way to go. To witness, in terms of absolute error the use of DCI+PACC allows to cut down quantification error to 3.3% on average, a substantial improvement with respect to the 9.2% on average obtained by just using DCI with a "classify and count" approach.

The combination of transfer learning (of which cross-lingual transfer is an instance) with quantification is an interesting task in general, that should prompt a body of dedicated research. We believe end-to-end approaches for cross-lingual quantification, not necessarily relying on classification as an intermediate step, would be worth exploring. Likewise, a natural extension of this work would be to explore applications of transfer learning to sentiment quantification different from the cross-lingual one, such as cross-domain sentiment quantification. Note also that, while this paper concentrates on a very narrow aspect of sentiment analysis (namely, Positive-Negative polarity detection), approaches such as the ones championed here can be in principle extended to deal with other labeling tasks in sentiment analysis, such as finer-grained polarity detection (e.g., using ordinal scales) or joint topic-sentiment detection.

ACKNOWLEDGEMENTS

The present work has been supported by the SoBigdata++ project, funded by the European Commission (Grant No. 871042) under Programme H2020-INFRAIA-2019-1. The authors opinions do not necessarily reflect those of the European Commission.

REFERENCES

1. P. Prettenhofer and B. Stein, "Cross-lingual adaptation using structural correspondence learning," *ACM Transactions on Intelligent Systems and Technology*, vol. 3, no. 1, 2011, p. Article 13, doi:10.1145/2036264.2036277.
2. A. Moreo, A. Esuli, and F. Sebastiani, "Distributional correspondence indexing for cross-lingual and cross-domain sentiment classification," *Journal of Artificial Intelligence Research*, vol. 55, 2016, pp. 131–163, doi: 10.1613/jair.4762.
3. P. González et al., "A review on quantification learning," *ACM Computing Surveys*, vol. 50, no. 5, 2017, pp. 74:1–74:40, doi:10.1145/3117807.
4. M. Ebrahimi, A. H. Yazdavar, and A. P. Sheth, "Challenges of Sentiment Analysis for Dynamic Events," *IEEE Intelligent Systems*, vol. 32, no. 5, 2017, pp. 70–75, doi: 10.1109/MIS.2017.3711649.
5. A. Esuli and F. Sebastiani, "Sentiment quantification," *IEEE Intelligent Systems*, vol. 25, no. 4, 2010, pp. 72–75.
6. M. S. Akhtar, A. Ekbal, and E. Cambria, "How Intense Are You? Predicting Intensities of Emotions and Sentiments using Stacked Ensemble," *IEEE Computational Intelligence Magazine*, vol. 15, no. 1, 2020, pp. 64–75.
7. W. Gao and F. Sebastiani, "From classification to quantification in tweet sentiment analysis," *Social Network Analysis and Mining*, vol. 6, no. 19, 2016, pp. 1–22, doi: 10.1007/s13278-016-0327-z.
8. K. Dashtipour et al., "Multilingual Sentiment Analysis: State of the Art and Independent Comparison of Techniques," *Cognitive Computation*, vol. 8, no. 4, 2016, pp. 757–771.
9. J. Barranquero, J. Díez, and J. J. del Coz, "Quantification-oriented learning based on reliable classifiers," *Pattern Recognition*, vol. 48, no. 2, 2015, pp. 591–604, doi:10.1016/j.patcog.2014.07.032.
10. A. Bella et al., "Quantification via probability estimators," *Proceedings of the 11th IEEE International Conference on Data Mining (ICDM 2010)*, 2010, pp. 737–742, doi: 10.1109/icdm.2010.75.
11. A. Esuli, A. Moreo, and F. Sebastiani, "A recurrent neural network for sentiment quantification," *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM 2018)*, 2018, pp. 1775–1778, doi:10.1145/3269206.3269287.
12. G. Forman, "Quantifying counts and costs via classification," *Data Mining and Knowledge Discovery*, vol. 17, no. 2, 2008, pp. 164–206, doi:10.1007/s10618-008-0097-y.
13. A. Moreo, A. Esuli, and F. Sebastiani, "Revisiting distributional correspondence indexing: A Python reimplementation and new experiments," arXiv:1810.09311 [cs.CL], 2018.
14. J. Blitzer, M. Dredze, and F. Pereira, "Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification," *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007)*, 2007, pp. 440–447.
15. H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society, Series B*, vol. 67, no. 2, 2005, pp. 301–320, doi:https://doi.org/10.1111/j.1467-9868.2005.00503.x.
16. F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, 2011, pp. 2825–2830.
17. B. Zadrozny and C. Elkan, "Transforming classifier scores into accurate multiclass probability estimates," *Proceedings of the 8th ACM International Conference on Knowledge Discovery and Data Mining (KDD 2002)*, 2002, pp. 694–699, doi:10.1145/775107.775151.
18. P. M. Domingos and M. J. Pazzani, "On the optimality of the simple Bayesian classifier under zero-one loss," *Machine Learning*, vol. 29, no. 2-3, 1997, pp. 103–130.
19. J. C. Platt, "Probabilistic outputs for support vector machines and comparison to regularized likelihood methods," A. Smola et al., eds., Advances in Large Margin ClassifiersThe MIT Press, Cambridge, MA, 2000, pp. 61–74.
20. F. Sebastiani, "Evaluation measures for quantification: An axiomatic approach," *Information Retrieval Journal*, 2019, doi:10.1007/s10791-019-09363-y.

Andrea Esuli is a tenured Researcher at the Italian National Research Council. His research interests include Machine Learning as applied to Text Mining, with a special emphasis on quantification and deep learning. Esuli has a PhD in Information Engineering from the University of Pisa. Contact him at andrea.esuli@isti.cnr.it.

Alejandro Moreo is a tenured Researcher at the Italian National Research Council. His research interests lie in Machine Learning and Text Mining, with an emphasis on transfer learning, cross-linguality, and representation learning. Moreo has a PhD in Computer Science from the University of Granada. Contact him at alejandro.moreo@isti.cnr.it.

Fabrizio Sebastiani is a tenured Director of Research at the Italian National Research Council. His research interests are in Machine Learning as applied to Text Mining, and especially in quantification and cost-sensitive learning. Contact him at fabrizio.sebastiani@isti.cnr.it.