SMILE: Twitter Emotion Classification using Domain Adaptation

Bo Wang **Maria Liakata** **Arkaitz Zubiaga** **Rob Procter** **Eric Jensen**
Department of Computer Science
University of Warwick
Coventry, UK
{bo.wang, m.liakata, e.jensen}@warwick.ac.uk

Abstract

Despite the widely spread research interest in social media sentiment analysis, sentiment and emotion classification across different domains and on Twitter data remains a challenging task. Here we set out to find an effective approach for tackling a cross-domain emotion classification task on a set of Twitter data involving social media discourse around arts and cultural experiences, in the context of museums. While most existing work in domain adaptation has focused on feature-based or/and instance-based adaptation methods, in this work we study a model-based adaptive SVM approach as we believe its flexibility and efficiency is more suitable for the task at hand. We conduct a series of experiments and compare our system with a set of baseline methods. Our results not only show a superior performance in terms of accuracy and computational efficiency compared to the baselines, but also shed light on how different ratios of labelled target-domain data used for adaptation can affect classification performance.

1 Introduction

With the advent and growth of social media as a ubiquitous platform, people increasingly discuss and express opinions and emotions towards all kinds of topics and targets. One of the topics that has been relatively unexplored in the scientific community is that of emotions expressed towards arts and cultural experiences. A survey conducted in 2012 by the British TATE Art Galleries found that 26 percent of the respondents had posted some kind of content online, such as blog posts, tweets or photos, about their experience in the art galleries during or after their visit [Villaespesa, 2013]. When cultural tourists share information about their experience in social media, this real-time communication and spontaneous engagement with art and culture not only broadens its target audience but also provides a new space where valuable insight shared by its customers can be garnered. As a result museums, galleries and other cultural venues have embraced social media such as Twitter, and actively used it to promote their exhibitions, organise participatory projects and/or create initiatives to engage with visitors, collecting valuable opinions and feedback (e.g. museum tweetups). This gold mine of user opinions has sparked an increasing research interest in the interdisciplinary field of social media and museum study [Fletcher and Lee, 2012; Villaespesa, 2013; Drotner and Schrøder, 2014].

We have also seen a surge of research in sentiment analysis with over 7,000 articles written on the topic [Feldman, 2013], for applications ranging from analyses of movie reviews [Pang and Lee, 2008] and stock market trends [Bollen *et al.*, 2011] to forecasting election results [Tumasjan *et al.*, 2010]. Supervised learning algorithms that require labelled training data have been successfully used for in-domain sentiment classification. However, cross-domain sentiment analysis has been explored to a much lesser extent. For instance, the phrase "light-weight" carries positive sentiment when describing a laptop but quite the opposite when it is used to refer to politicians. In such cases, a classifier trained on one domain may not work well on other domains. A widely adopted solution to this problem is domain adaptation, which allows building models from a fixed set of source domains and deploy them into a different target domain. Recent developments in sentiment analysis using domain adaptation are mostly based on feature-representation adaptation [Blitzer *et al.*, 2007; Pan *et al.*, 2010; Bollegala *et al.*, 2011], instance-weight adaptation [Jiang and Zhai, 2007; Xia *et al.*, 2014; Tsakalidis *et al.*, 2014] or combinations of both [Xia *et al.*, 2013; Liu *et al.*, 2013]. Despite its recent increase in popularity, the use of domain adaptation for sentiment and emotion classification across topics on Twitter is still largely unexplored [Liu *et al.*, 2013; Tsakalidis *et al.*, 2014; Townsend *et al.*, 2014].

In this work we set out to find an effective approach for tackling the cross-domain emotion classification task on Twitter, while also furthering research in the interdisciplinary study of social media discourse around arts and cultural experiences[1]. We investigate a model-based adaptive-SVM approach that was previously used for video concept detection [Yang *et al.*, 2007] and compare with a set of domain-dependent and domain-independent strategies. Such a model-based approach allows us to directly adapt existing models to the new target-domain data without having to generate domain-dependent features or adjusting weights for each of

[1] SMILE project: `http://www.culturesmile.org/`

Proceedings of the 4th Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2016), IJCAI 2016, pages 15-21, New York City, USA, July 10, 2016.

the training instances.We conduct a series of experiments and evaluate the proposed system[2] on a set of Twitter data about museums, annotated by three annotators from the social sciences. The aim is to maximise the use of the base classifiers that were trained from a general-domain corpus, and through domain adaptation minimise the classification error rate across 5 emotion categories: *anger*, *disgust*, *happiness*, *surprise* and *sadness*. Our results show that adapted SVM classifiers achieve significantly better performance than out-of-domain classifiers and also suggest a competitive performance compared to in-domain classifiers. To the best of our knowledge this is the first attempt at cross-domain emotion classification for Twitter data.

2 Related Work

Most existing approaches can be classified into two categories: feature-based adaptation and instance-based adaptation. The former seek to construct new adaptive feature representations that reduce the difference between domains, while the latter aims to sample and re-weight source domain training data for use in classification within the target domain.

With respect to feature domain adaptation, [Blitzer *et al.*, 2007] applied structural correspondence learning (SCL) algorithm for cross-domain sentiment classification. SCL chooses a set of *pivot features* with highest mutual information to the domain labels, and uses these pivot features to align other features by training N linear predictors. Finally it computes singular value decomposition (SVD) to construct low-dimensional features to improve its classification performance. A small amount of target domain labelled data is used to learn to deal with misaligned features from SCL. [Townsend *et al.*, 2014] found that SCL did not work well for cross-domain adaptation of sentiment on Twitter due to the lack of mutual information across the Twitter domains and uses subjective proportions as a backoff adaptation approach. [Pan *et al.*, 2010] proposed to construct a bipartite graph from a co-occurrence matrix between domain-independent and domain specific features to reduce the gap between different domains and use spectral clustering for feature alignment. The resulting clusters are used to represent data examples and train sentiment classifiers. They used mutual information between features and domains to classify domain-independent and domain specific features, but in practice this also introduces mis-classification errors. [Bollegala *et al.*, 2011] describes a cross-domain sentiment classification approach using an automatically created sentiment sensitive thesaurus. Such a thesaurus is constructed by computing the point-wise mutual information between a lexical element u and a feature as well as relatedness between two lexical elements. The problem with these feature adaptation approaches is that they try to connect domain-dependent features to known or common features under the assumption that parallel sentiment words exist in different domains, which is not necessarily applicable to various topics in tweets [Liu *et al.*, 2013]. [Glorot *et al.*, 2011] proposes a deep learning system to extract features that are highly beneficial for the domain adaptation of sentiment classifiers, under the intuition that deep learning algorithms learn intermediate concepts (between raw input and target) and these intermediate concepts could yield better transfer across domains.

When it comes to instance adaptation, [Jiang and Zhai, 2007] proposes an instance weighting framework that prunes "misleading" instances and approximates the distribution of instances in the target domain. Their experiments show that by adding some labelled target domain instances and assigning higher weights to them performs better than either removing "misleading" source domain instances using a small number of labelled target domain data or bootstrapping unlabelled target instances. [Xia *et al.*, 2014] adapts the source domain training data to the target domain based on a logistic approximation. [Tsakalidis *et al.*, 2014] learns different classifiers on different sets of features and combines them in an ensemble model. Such an ensemble model is then applied to part of the target domain test data to create new training data (i.e. documents for which different classifiers had the same predictions). We include this ensemble method as one of our baseline approaches for evaluation and comparison.

In contrast with most cross-domain sentiment classification works, we use a model-based approach proposed in [Yang *et al.*, 2007], which directly adapts existing classifiers trained on general-domain corpora. We believe this is more efficient and flexible [Yang and Hauptmann, 2008] for our task. We evaluate on a set of manually annotated tweets about cultural experiences in museums and conduct a finer-grained classification of emotions conveyed (i.e. *anger*, *disgust*, *happiness*, *surprise* and *sadness*).

3 Datasets

We use two datasets, a source-domain dataset and a target-domain dataset, which enables us to experiment on domain adaptation. The source-domain dataset we adopted is the general-domain Twitter corpus created by [Purver and Battersby, 2012], which was generated through distant supervision using hashtags and emoticons associated with 6 emotions: anger, disgust, fear, happiness, surprise and sadness.

Our target-domain dataset that allows us to perform experiments on emotions associated with cultural experiences consists of a set of tweets pertaining to museums. A collection of tweets mentioning one of the following Twitter handles associated with British museums was gathered between May 2013 and June 2015: @camunivmuseums, @fitzmuseum_uk, @kettlesyard, @maacambridge, @iciabath, @thelmahulbert, @rammuseum, @plymouthmuseum, @tateliverpool, @tate_stives, @nationalgallery, @britishmuseum, @_thewhitechapel. These are all museums associated with the SMILES project. A subset of 3,759 tweets was sampled from this collection for manual annotation. We developed a tool for manual annotation of the emotion expressed in each of these tweets. The options for the annotation of each tweet included 6 different emotions; the six Ekman emotions as in [Purver and Battersby, 2012], with the exception of 'fear' as it never featured in the context of tweets about museums. Two extra annotation options were included to indicate that a tweet should have *no code*, indicating that a tweet was

[2] The code can be found at `http://bit.ly/1WHup4b`

not conveying any emotions, and *not relevant* when it did not refer to any aspects related to the museum in question. The annotator could choose more than one emotion for a tweet, except when *no code* or *not relevant* were selected, in which case no additional options could be picked. The annotation of all the tweets was performed independently by three sociology PhD students. Out of the 3,759 tweets that were released for annotation, at least 2 of the annotators agreed in 3,085 cases (82.1%). We use the collection resulting from these 3,085 tweets as our target-domain dataset for classifier adaptation and evaluation. Note that tweets labelled as *no code* or *not relevant* are included in our dataset to reflect a more realistic data distribution on Twitter, while our source-domain data doesn't have any *no code* or *not relevant* tweets.

The distribution of emotion annotations in Table 2 shows a remarkable class imbalance, where *happy* accounts for 30.2% of the tweets, while the other emotions are seldom observed in the museum dataset. There is also a large number of tweets with no emotion associated (41.8%). One intuitive explanation is that Twitter users tend to express positive and appreciative emotions regarding their museum experiences and shy away from making negative comments. This can also be demonstrated by comparing the museum data emotion distribution to our general-domain source data as seen in Figure 1, where the sample ratio of positive instances is shown for each emotion category.

To quantify the difference between two text datasets, Kullback-Leibler (KL) divergence has been commonly used before [Dai *et al.*, 2007]. Here we use the KL-divergence method proposed by [Bigi, 2003], as it suggests a back-off smoothing method that deals with the data sparseness problem. Such back-off method keeps the probability distributions summing to 1 and allows operating on the entire vocabulary, by introducing a normalisation coefficient and a very small threshold probability for all the terms that are not in the given vocabulary. Since our source-domain data contains many more tweets than the target-domain data, we have randomly sub-sampled the former and made sure the two data sets have similar vocabulary size in order to avoid biases. We removed stop words, user mentions, URL links and re-tweet symbols prior to computing the KL-divergence. Finally we randomly split each data set into 10 folds and compute the in-domain and cross-domain symmetric KL-divergence (KLD) value between every pair of folds. Table 1 shows the computed KL-divergence averages. It can be seen that KL-divergence between the two data sets (i.e. KLD($D_{src} \,||\, D_{tar}$)) is twice as large as the in-domain KL-divergence values. This suggests a significant difference between data distributions in the two domain and thus justifies our need for domain adaptation.

Data domain	**Averaged KLD value**
KLD($D_{src} \,\|\|\, D_{src}$)	2.391
KLD($D_{tar} \,\|\|\, D_{tar}$)	2.165
KLD($D_{src} \,\|\|\, D_{tar}$)	4.818

Table 1: In-domain and cross-domain KL-divergence values

Emotion	**No. of tweets**	**% of tweets**
no code	1572	41.8%
happy	1137	30.2%
not relevant	214	5.7%
anger	57	1.5%
surprise	35	0.9%
sad	32	0.9%
happy & surprise	11	0.3%
happy & sad	9	0.2%
disgust & anger	7	0.2%
disgust	6	0.2%
sad & anger	2	0.1%
sad & disgust	2	0.1%
sad & disgust & anger	1	<0.1%

Table 2: Target data emotion distribution

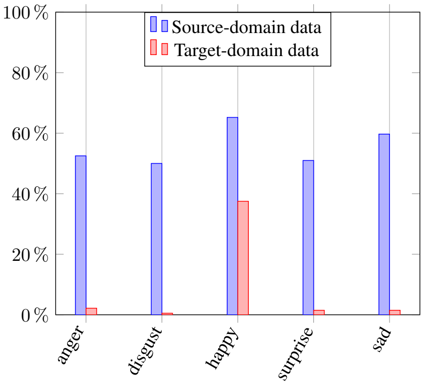


Figure 1: Source and target data distribution comparison

4 Methodology

Given the source-domain D_{src} and target-domain D_{tar}, we have one or k sets of labelled source-domain data denoted as $\left\{(x_i^k, y_i^k)\right\}_{i=1}^{N_{src}^k}$ in D_{src}, where x_i^k is the i_{th} feature vector with each element as the value of the corresponding feature and y_i^k are the emotion categories that the i_{th} instance belongs to. Suppose we have some classifiers $f_{src}^k(x)$ that have been trained on the source-domain data (named as the *auxiliary classifiers* in [Yang *et al.*, 2007]) and a small set of labelled target-domain data as D_{tar}^l where $D_{tar} = D_{tar}^l \cup D_{tar}^u$, our goal is to adapt $f_{src}^k(x)$ to a new classifier $f_{tar}(x)$ based on the small set of labelled examples in D_{tar}^l, so it can be used to accurately predict the emotion class of unseen data from D_{tar}^u.

4.1 Base Classifiers

Our base classifiers are the classifiers that have been trained on the source-domain data $\left\{(x_i, y_i)\right\}_{i=1}^{N_{src}}$, where $y_i \in \{1, ..., K\}$ with K referring to the number of emotion categories. In our work, we use Support Vector Machines (SVMs) in a "one-versus-all" setting, which trains K binary classifiers, each separating one class from the rest. We chose this as a better way of dealing with class imbalance in a multi-class scenario.

Features

The base classifiers are trained on 3 sets of features generated from the source-domain data: (i) n-grams, (ii) lexicon features, (iii) word embedding features.

N-gram models have long been used in NLP for various tasks. We used 1-2-3 grams after filtering out all the stop words, as our n-gram features. We construct 32 **Lexicon features** from 9 Twitter specific and general-purpose lexica. Each lexicon provides either a numeric sentiment score, or categories where a category could correspond to a particular emotion or a strong/weak positive/negative sentiment.

The use of **Word embedding features** to represent the context of words and concepts, has been shown to be very effective in boosting the performance of sentiment classification. In this work we use a set of word embeddings learnt using a sentiment-specific method in [Tang *et al.*, 2014] and another set of general word embeddings trained with 5 million tweets by [Vo and Zhang, 2015]. Training on an additional set of 3 million tweets we trained ourselves did not increase performance. Pooling functions are essential and particularly effective for feature selection from dense embedding feature vectors. [Tang *et al.*, 2014] applied the *max*, *min* and *mean* pooling functions and found them to be highly useful. We tested and evaluated six pooling functions, namely *sum*, *max*, *min*, *mean*, *std* (i.e. standard deviation) and *product*, and selected *sum*, *max* and *mean* as they led to the best performance.

4.2 Classifier Adaptation

[Yang *et al.*, 2007] proposes a many-to-one SVM adaptation model, which directly modifies the decision function of an ensemble of existing classifiers $f^k_{src}(x)$, trained with one or k sets of labelled source-domain data in D_{src}, and thus creates a new adapted classifier $f_{tar}(x)$ for the target-domain D_{tar}. The adapted classifier has the following form:

$$f_{tar}(x) = \sum_{k=1}^{M} \tau^k f^k_{src}(x) + \Delta f(x) \tag{1}$$

where $\tau^k \in (0, 1)$ is the weight of each base classifier $f^k_{src}(x)$. $\Delta f(x)$ is the perturbation function that is learnt from a small set of labelled target-domain data in D^l_{tar}. As shown in [Yang *et al.*, 2007] it has the form:

$$\Delta f(x) = w^T \phi(x) = \sum_{i=1}^{N} \alpha_i y_i \boldsymbol{K}(x_i, x) \tag{2}$$

where $w = \sum_{i=1}^{N} \alpha_i y_i \phi(x_i)$ are the model parameters to be estimated from the labelled examples in D^l_{tar} and α_i is the feature coefficient of the i_{th} labelled target-domain instance. Furthermore $\boldsymbol{K}(\cdot, \cdot) \equiv \phi(\cdot)^T \phi(\cdot)$ is the kernel function induced from the nonlinear feature mapping. $\Delta f(x)$ is learnt in a framework that aims to minimise the regularised empirical risk [Yang, 2009]. The adapted classifier $f_{tar}(x)$ learnt under this framework tries to minimise the classification error on the labelled target-domain examples and the distance from the base classifiers $f^k_{src}(x)$, to achieve a better bias-variance trade-off.

In this work we use the extended multi-classifier adaptation framework proposed by [Yang and Hauptmann, 2008], which allows the weight controls $\{\tau^k\}^M_{k=1}$ of the base classifiers $f^k_{src}(x)$ to be learnt automatically based on their classification performance of the small set of labelled target-domain examples. To achieve this, [Yang and Hauptmann, 2008] adds another regulariser to the regularised loss minimisation framework, with the objective function of training the adaptive classifier now written as:

$$\begin{aligned} \min_{w,\tau,\xi} \quad & \frac{1}{2} w^T w + \frac{1}{2} B(\tau)^T \tau + C \sum_{i=1}^{N} \xi_i \\ \text{s.t.} \quad & y_i \sum_{k=1}^{M} \tau^k f^k_{src}(x) + y_i w^T \phi(x_i) \geq 1 - \xi_i, \\ & \xi^m_i \geq 0, \forall (x_i, y_i) \in D_{src} \end{aligned} \tag{3}$$

where $\frac{1}{2}(\tau)^T \tau$ measures the overall contribution of base classifiers. Thus this objective function seeks to avoid over reliance on the base classifiers and also over-complex $\Delta f(\cdot)$. The two goals are balanced by the parameter B. By rewriting this objective function as a minimisation problem of a Lagrange (primal) function and set its derivative against w, τ, and ξ to zero, we have:

$$w = \sum_{i=1}^{N} \alpha_i y_i \phi(x_i), \quad \tau^k = \frac{1}{B} \sum_{i=1}^{N} \alpha_i y_i f^k_{src}(x_i) \tag{4}$$

where τ^k is a weighted sum of $y_i f^k_{src}(x_i)$ and it indicates the classification performance of f^k_{src} on the target-domain. Therefore we have base classifiers assigned with larger weight if they classify the labelled target-domain data well. Now given (1), (2) and (4), the new decision function can be formulated as:

$$\begin{aligned} f_{tar}(x) &= \frac{1}{B} \sum_{k=1}^{M} \sum_{i=1}^{N} \alpha_i y_i f^k_{src}(x_i) f^k_{src}(x) + \Delta f(x) \\ &= \sum_{i=1}^{N} \alpha_i y_i \Big(\boldsymbol{K}(x_i, x) + \frac{1}{B} \sum_{k=1}^{M} f^k_{src}(x_i) f^k_{src}(x) \Big) \end{aligned} \tag{5}$$

Comparing (5) with a standard SVM model $f(x) = \sum_{i=1} \alpha_i y_i \boldsymbol{K}(x_i, x)$, this multi-classifier adaptation model can be interpreted as a way of adding the predicted labels of base classifiers on the target-domain as additional features. Under this interpretation the scalar B balances the contribution of the original features and additional features.

4.3 Data Preprocessing

A set of preprocessing techniques applied include substituting URL links with strings “URL”, user mentions with “@USERID”, removing the hashtag symbol “#”, normalising emoticons and abbreviations[3].

5 Results and Evaluation

In this section we present the experimental results and compare our proposed adaptation system with a set of domain-dependent and domain-independent strategies. We also investigate the effect of different sizes of the labelled target-domain data in the classification performance.

[3] http://bit.ly/1U7fiQR

5.1 Adaptation Baselines

The baseline methods and our proposed system are the following:

- **BASE**: the base classifiers use either one set of features or all three feature sets (i.e. BASE-all). As an example, the BASE-embedding classifier is trained and tuned with all source-domain data using only word-embedding features, then tested on 30% of our target-domain data. We use the LIBSVM implementation [Chang and Lin, 2011] of SVM for building the base classifiers.
- **TARG**: trained and tuned with 70% labelled target-domain data. Since this model is entirely trained from the target domain, it can be considered as the *performance upper-bound* that is very hard to beat.
- **AGGR**: an aggregate model trained from all source-domain data and 70% labelled target-domain data.
- **ENSEMBLE**: combines the base classifiers in an ensemble model. Then perform classification on 30% of the target-domain data to generate new training data, as described in Section 2.
- **ADAPT**: our domain adapted models using either one base classifier trained with all feature sets (i.e. ADAPT-1-model) or an ensemble of three standalone base classifiers with each trained with one set of features (i.e. ADAPT-3-model). We use 30% of the labelled target-domain data for classifier adaptation and parameter tuning described in Section 4.2.

The above methods are all tested on the same 30% labelled target-domain data in order to make their results comparable. In addition we perform in-domain cross-validation and evaluation only on our source-domain data using all feature sets; this model is named as SRC-all. We use an RBF kernel function (as it outperforms linear kernel. Polynomial kernel gives similar performance but requires more parameter tuning) with default setting of the gamma parameter γ in all the methods. For the cost factor C and class weight parameter (except the SRC-all model) we conduct cross-validated grid-search over the same set of parameter values for all the methods, for parameter optimisation. This makes sure our ADAPT models are comparable with BASE, TARG, ENSEMBLE and AGGR. For ADAPT-3-model we also optimise the base classifier weight parameters, denoted as τ^k in Eq.(1), as described in Section 4.2.

5.2 Experimental Results

We report the experimental results in **Table 3**, with three categories of models: 1) in-domain no adaptation methods, i.e. BASE and TARG models, TARG being the *upper-bound* for performance evaluation; 2) the domain adaptation baselines, i.e. AGGR and ENSEMBLE and 3) our adaptation systems (ADAPT models). As can be seen the classification performances reported for emotions other than "happy" are below 50 in terms of $\mathbf{F_1}$ score with some results being as low as 0.00. This is caused by the class imbalance issue within these emotions as shown in Table 2 and Figure 1, especially for the emotion "disgust" which has only 16 tweets. We tried to balance this issue using a class weight parameter, but it still is very challenging to overcome without acquiring more labelled data than we currently have. It especially effects our domain adaptation as all the parameters in Eq.(3) cannot be properly optimised.

Since there are very few tweets annotated as "disgust", we decide not to consider the "disgust" emotion as part of our experiment evaluation here. As seen in Table 3, BASE models are outperformed significantly by all other methods (except ENSEMBLE, which performs only slightly better than the BASE models) positing the importance of domain adaptation. With the exception of the ADAPT-3-model for "Anger", our ADAPT models consistently outperform AGGR-all and ENSEMBLE while showing competitive performance compared to the *upper-bound* baseline, TARG-all. We also observe that the aggregation model AGGR-all is outperformed by TARG-all, indicating such domain knowledge cannot be transferred effectively to a different domain by simply modelling from aggregated data from both domains. In comparison, our ADAPT models are able to leverage the large and balanced source-domain data (as base classifiers) unlike TARG, while adjusting the contribution of each base classifier unlike AGGR.

When comparing our ADAPT models, we find that in most cases models adapted from multiple base classifiers beat the ones adapted from one single base classifier, even though the same features are used in both scenarios. This shows the benefit of the multi-classifier adaptation approach, which aims to maximise the utility of each base classifier. Two additional models, namely ADAPT-1-modelx and ADAPT-3-modelx, are the replicates of ADAPT-1/3-model except they also use 40% target-domain data for tuning the model parameters. On average their results are only slightly better than ADAPT-1/3-model that use 30% of the target-domain data for both training and parameter optimisation. This is especially prominent with "happiness" where we have sufficient target-domain instances and less of a class imbalance issue. This shows our ADAPT models are able to yield knowledge transfer effectively across different domains with a small amount of labelled target-domain data. More analysis on the impact of adaptation sample ratios is given in Section 5.3.

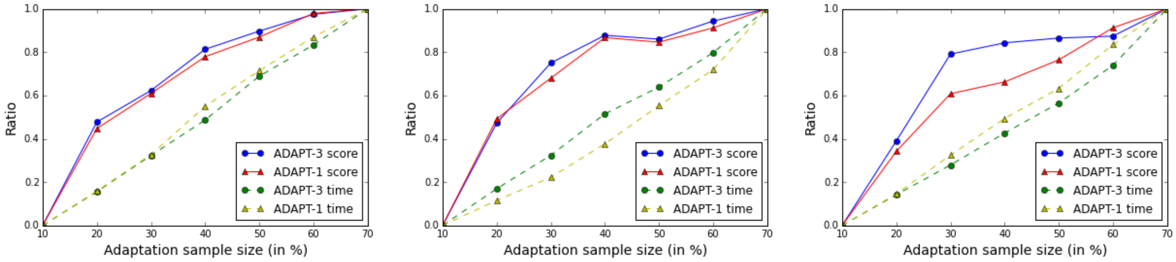
We can also evaluate the performance of each model by comparing its efficiency in terms of computation time. Here we report the total computation time taken for all the above methods except BASE, for the emotion "happiness". Such computation process consists of adaptation training, grid-search over the same set of parameter values and final testing. As seen in Table 4, compared to other out-of-domain strategies the proposed ADAPT models are more efficient to train especially in comparison with AGGR, which is an order of magnitude more costly due to the inclusion of source-domain data. Within the ADAPT models, ADAPT-1-model requires less time to train since it only has one base classifier for adaptation.

5.3 Effect of Adaptation Training Sample ratios

Here we evaluate the effect of different ratios of the labelled target-domain data on the overall classification performance for the emotion "happiness". Figure 2 shows the normalised $\mathbf{F_1}$ scores and computation time of each ADAPT

Model	Anger			Disgust			Happy			Surprise			Sad		
	P	**R**	$\mathbf{F_1}$	**P**	**R**	$\mathbf{F_1}$	**P**	**R**	$\mathbf{F_1}$	**P**	**R**	$\mathbf{F_1}$	**P**	**R**	$\mathbf{F_1}$
BASE-ngrams	5.77	40.91	10.11	0.49	100.0	0.97	37.62	100.0	54.67	1.46	100.0	2.87	1.50	100.0	2.96
BASE-lexicon	2.59	90.91	5.03	0.55	100.0	1.10	38.43	98.96	55.36	0.00	0.00	0.00	2.54	93.33	4.94
BASE-embedding	2.06	72.73	4.02	0.00	0.00	0.00	39.18	96.11	55.66	2.00	60.00	3.88	1.49	80.00	2.92
BASE-all	2.01	59.09	3.88	5.00	20.00	8.00	38.75	98.19	55.57	1.69	66.67	3.29	1.58	86.67	3.11
TARG-all	36.00	40.91	**38.30**	0.00	0.00	0.00	78.04	84.72	81.24	20.83	33.33	**25.64**	18.75	20.00	**19.35**
AGGR-all	10.71	27.27	15.38	33.33	20.00	25.00	64.79	86.27	74.00	5.88	11.11	7.69	4.17	20.00	6.90
ENSEMBLE	2.11	100.0	4.13	0.49	100.0	0.97	45.20	83.55	58.66	2.70	11.11	4.35	1.46	100.0	2.88
ADAPT-1-model	16.28	31.82	21.54	0.59	80.00	1.18	79.34	80.57	79.95	11.11	13.33	12.12	100.0	6.67	12.50
ADAPT-3-model	20.00	9.09	12.50	0.00	0.00	0.00	82.11	80.83	81.46	8.14	46.67	13.86	8.77	33.33	13.89
ADAPT-1-modelx	21.43	13.64	16.67	100.0	20.00	**33.33**	80.53	79.27	79.90	12.50	26.67	17.02	16.67	13.33	14.81
ADAPT-3-modelx	20.00	22.73	21.28	1.82	20.00	3.33	80.30	83.42	**81.83**	12.50	26.67	17.02	10.20	33.33	15.63
SRC-all	93.57	93.37	93.46	99.05	98.73	98.89	81.87	85.91	83.83	96.25	98.03	97.13	91.04	92.51	91.76

Table 3: Model performance comparison



(a) C = 1

(b) C = 3

(c) C = 10

Figure 2: Performance of each ADAPT model with C = 1,3,10 vs. its computation time

Model	Total computation time in minutes
TARG-all	7.72
ENSEMBLE	209.72
AGGR-all	1238.24
ADAPT-1-model	26.30
ADAPT-3-model	118.41

Table 4: Total computation time for each method

model across different adaptation training sample sizes ranging from 10% to 70% of the total target-domain data (with the same 30% held out as test data) and with the cost factor C = 1, 3 and 10 (as the same choices of C are used in [Yang *et al.*, 2007] for conducting their experiment). We observe a logarithmic growth for the $\mathbf{F_1}$ scores obtained from every model, against a linear growth of computation time cost. Thus even though there is a reasonable increase in classification performance when increasing the adaptation sample size from 50% to 70%, it becomes much less efficient to train such models and we require more data, which may not be available. Since we have a trade-off between model effectiveness and efficiency here, it is appropriate to use 30% of our labelled target-domain data for classifier adaptation as we have done so in ADAPT-1-model and ADAPT-3-model. One should select the adaptation training sample size accordingly based on the test data at hand, but empirically we think 1,000 labelled target-domain tweets would be enough for an effective adaptation to classify 3,000-4,000 test tweets.

6 Conclusion

In this work we study a model-based multi-class adaptive-SVM approach to cross-domain emotion recognition and compare against a set of domain-dependent and domain-independent strategies. We conduct a series of experiments and evaluate our proposed system on a set of newly annotated Twitter data about museums. We find that our adapted SVM model outperforms the out-of-domain base models and domain adaptation baselines while also showing competitive performance against the in-domain *upper-bound* model. Moreover, in comparison to other adaptation strategies our approach is computationally more efficient especially compared to the classifier trained on aggregated source and target data. Finally, we shed light on how different ratios of labelled target-domain data used for adaptation can effect classification performance. We show there is a trade-off between model effectiveness and efficiency when selecting adaptation sample size. Our code and data[4] are publicly available, enabling further research and comparison with our approach.

In the future we would like to investigate a feature-based deep learning approach for cross-topic emotion classification on Twitter while examining the possibility of making it as efficient and flexible as the model adaptation based approaches. Another future direction is to study how to best resolve the remarkable class imbalance issue in social media emotion analysis when some emotions are rarely expressed.

[4] http://bit.ly/1SddvIw

Acknowledgments

This work has been funded by the AHRC SMILES project. We would like to thank Liz Walker, Matt Jeffryes and Michael Clapham for their contribution to earlier versions of the emotion classifiers.

References

[Bigi, 2003] Brigitte Bigi. *Using Kullback-Leibler distance for text categorization*. Springer, 2003.

[Blitzer *et al.*, 2007] John Blitzer, Mark Dredze, Fernando Pereira, et al. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *ACL*, volume 7, pages 440–447, 2007.

[Bollegala *et al.*, 2011] Danushka Bollegala, David Weir, and John Carroll. Using multiple sources to construct a sentiment sensitive thesaurus for cross-domain sentiment classification. In *NAACL HLT*, pages 132–141. Association for Computational Linguistics, 2011.

[Bollen *et al.*, 2011] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.

[Chang and Lin, 2011] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

[Dai *et al.*, 2007] Wenyuan Dai, Gui-Rong Xue, Qiang Yang, and Yong Yu. Co-clustering based classification for out-of-domain documents. In *SIGKDD*, pages 210–219. ACM, 2007.

[Drotner and Schrøder, 2014] Kirsten Drotner and Kim Christian Schrøder. *Museum communication and social media: The connected museum*. Routledge, 2014.

[Feldman, 2013] Ronen Feldman. Techniques and applications for sentiment analysis. *Communications of the ACM*, 56(4):82–89, 2013.

[Fletcher and Lee, 2012] Adrienne Fletcher and Moon J Lee. Current social media uses and evaluations in american museums. *Museum Management and Curatorship*, 27(5):505–521, 2012.

[Glorot *et al.*, 2011] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *ICML*, pages 513–520, 2011.

[Jiang and Zhai, 2007] Jing Jiang and ChengXiang Zhai. Instance weighting for domain adaptation in nlp. In *ACL*, pages 264–271. Association for Computational Linguistics, June 2007.

[Liu *et al.*, 2013] Shenghua Liu, Fuxin Li, Fangtao Li, Xueqi Cheng, and Huawei Shen. Adaptive co-training svm for sentiment classification on tweets. In *CIKM*, pages 2079–2088. ACM, 2013.

[Pan *et al.*, 2010] Sinno Jialin Pan, Xiaochuan Ni, Jian-Tao Sun, Qiang Yang, and Zheng Chen. Cross-domain sentiment classification via spectral feature alignment. In *WWW*, pages 751–760. ACM, 2010.

[Pang and Lee, 2008] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135, 2008.

[Purver and Battersby, 2012] Matthew Purver and Stuart Battersby. Experimenting with distant supervision for emotion classification. In *EACL*, pages 482–491. Association for Computational Linguistics, 2012.

[Tang *et al.*, 2014] Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. Learning sentiment-specific word embedding for twitter sentiment classification. In *ACL*, volume 1, pages 1555–1565, 2014.

[Townsend *et al.*, 2014] Richard Townsend, Aaron Kalair, Ojas Kulkarni, Rob Procter, and Maria Liakata. University of warwick: Sentiadaptron-a domain adaptable sentiment analyser for tweets-meets semeval. *SemEval 2014*, page 768, 2014.

[Tsakalidis *et al.*, 2014] Adam Tsakalidis, Symeon Papadopoulos, and Ioannis Kompatsiaris. An ensemble model for cross-domain polarity classification on twitter. In *WISE*, pages 168–177. Springer, 2014.

[Tumasjan *et al.*, 2010] Andranik Tumasjan, Timm Oliver Sprenger, Philipp G Sandner, and Isabell M Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM*, 10:178–185, 2010.

[Villaespesa, 2013] Elena Villaespesa. Diving into the museums social media stream: Analysis of the visitor experience in 140 characters. In *Museums and the Web*, 2013.

[Vo and Zhang, 2015] Duy-Tin Vo and Yue Zhang. Target-dependent twitter sentiment classification with rich automatic features. In *IJCAI*, pages 1347–1353, 2015.

[Xia *et al.*, 2013] Rui Xia, Chengqing Zong, Xuelei Hu, and Erik Cambria. Feature ensemble plus sample selection: domain adaptation for sentiment classification. *Intelligent Systems, IEEE*, 28(3):10–18, 2013.

[Xia *et al.*, 2014] Rui Xia, Jianfei Yu, Feng Xu, and Shumei Wang. Instance-based domain adaptation in nlp via in-target-domain logistic approximation. In *AAAI*, 2014.

[Yang and Hauptmann, 2008] Jun Yang and Alexander G Hauptmann. A framework for classifier adaptation and its applications in concept detection. In *MIR*, pages 467–474. ACM, 2008.

[Yang *et al.*, 2007] Jun Yang, Rong Yan, and Alexander G Hauptmann. Cross-domain video concept detection using adaptive svms. In *Proceedings of the 15th international conference on Multimedia*, pages 188–197. ACM, 2007.

[Yang, 2009] Jun Yang. *A general framework for classifier adaptation and its applications in multimedia*. PhD thesis, Columbia University, 2009.