Algorithm 1: Cross-Domain Text Classification via Topic Correlation Analysis

```
Input  : (1) Source domain labeled data set $\mathcal{D}^s$ and
         target domain unlabeled data set $\mathcal{D}^t$, (2)
         Number of shared topics $K$, (3) Number of
         source/target domain-specific topics $K^s$, $K^t$,
         (4) Number of iterations $T$
Output : The predicted class label of each unlabeled
         document $d_i^t \in \mathcal{D}^t$ in the target domain
Initialize the parameters of the proposed JMM model;
for $t \leftarrow 1$ to $T$ do
    E-step: Compute the posterior probabilities with
    the E-step updates in Figure 1;
    M-step: Update the model parameters with the
    M-step updates in Figure 1 and (3);
end
Measure the topic correlations using (4) and (5);
Construct the topic mapping matrix $\mathbf{U}$ using (6);
Represent each document with extracted topics as (7);
Transform each document in the target domain as (8);
Train a classifier with the labeled documents
$\mathcal{D}^s = \{(\phi(d_n^s), y_n^s)\}_{n=1}^{N^s}$ and predict the class labels of
the unlabeled documents $\mathcal{D}^t = \{\psi(\phi(d_i^t))\}_{i=1}^{N^t}$.
```

text classification algorithms (Dai et al. 2007; Xue et al. 2008; Pan and Yang 2010). It contains nearly 20,000 newsgroup documents which have been evenly partitioned into 20 different newsgroups. As in the previous works (Dai et al. 2007; Xue et al. 2008), we generate six cross-domain text data sets from 20Newsgroups by utilizing its hierarchical structure. Specifically, the learning task is defined as the top-category binary classification, where our goal is to classify documents into one of the top-categories (e.g., *Comp*, *Rec*, etc.). For each data set, we select one top-category (e.g., *Comp*) as the positive class and another top-category (e.g., *Rec*) as the negative class. Then we select some subcategories (e.g., *comp.graphics* and *rec.motorcycles*) under the positive and the negative classes respectively to form a domain. In this work, we use the documents from four top-categories: *Comp*, *Rec*, *Sci* and *Talk* to generate data sets. Table 2 summarizes the data sets generated from 20Newsgroups.

Reuters-21578 The Reuters-21578 is another famous data set for evaluating text classification algorithms (Dai et al. 2007). As 20Newsgroups, the documents in Reuters-21578 are also organized with a hierarchical structure. For Reuters-21578, we use the preprocessed version of data sets provided in the web site (http://www.cse.ust.hk/TL/index.html) for experiments. This data set contains three cross-domain data sets which are generated with the documents from three biggest top-categories (i.e., *Orgs*, *People* and *Places*). Table 3 summarizes the generated data sets.

Baselines and Evaluation Criteria

To test the effectiveness of TCA, we compare it with two conventional classification algorithms: Support Vector Machine (SVM) and Logistic Regression (LG), and three state-of-the-art cross-domain classification methods: Spectral Feature Alignment (SFA) (Pan et al. 2010), Topic-bridge PLSA (TPLSA) (Xue et al. 2008) and Collaborative Dual-PLSA (CDPLSA) (Zhuang et al. 2010). For SVM and LG, the classifiers are trained with the labeled documents from the source domain and used to predict the class labels of unlabeled documents in the target domain. In SFA, the spectral clustering algorithm is adapted to co-cluster all words into the shared clusters for domain adaptation. Both TPLSA and CDPLSA aim to jointly model documents from different domains based on topic modeling. In TPLSA, all topics are assumed to be shared by different domains and used to represent documents. CDPLSA jointly models different domains by assuming that the associations between the topics and the document categories are stable across domains. In order to verify the usefulness of the induced feature mapping between domain-specific topics, we modify TCA by using only the shared topics to represent documents for classifier training. We denote it TCA$^{\text{share}}$. The classification accuracy is adopted as the evaluation criteria. For the algorithms which have the random initialization process, we conduct 10 and 50 random runs for the experiments on 20Newsgroups and Reuters-21578, respectively. And the average results of the random runs are reported.

Table 2: Data Sets Generated from 20Newsgroups

Data set	Source Domain $\mathcal{D}^s$	Target Domain $\mathcal{D}^t$
Comp vs Rec	comp.graphics	comp.os.ms-windows.misc
	comp.sys.ibm.pc.hardware	comp.sys.mac.hardware
	rec.motorcycles	rec.autos
	rec.sport.baseball	rec.sport.hockey
Comp vs Sci	comp.os.ms-windows.misc	comp.graphics
	comp.sys.ibm.pc.hardware	comp.sys.mac.hardware
	sci.electronics	sci.crypt
	sci.space	sci.med
Comp vs Talk	comp.os.ms-windows.misc	comp.graphics
	comp.sys.ibm.pc.hardware	comp.sys.mac.hardware
	talk.politics.mideast	talk.politics.guns
	talk.politics.misc	talk.religion.misc
Rec vs Sci	rec.autos	rec.motorcycles
	rec.sport.baseball	rec.sport.hockey
	sci.crypt	sci.electronics
	sci.med	sci.space
Rec vs Talk	rec.autos	rec.motorcycles
	rec.sport.baseball	rec.sport.hockey
	talk.politics.mideast	talk.politics.guns
	talk.politics.misc	talk.religion.misc
Sci vs Talk	sci.crypt	sci.electronics
	sci.med	sci.space
	talk.politics.misc	talk.politics.guns
	talk.religion.misc	talk.politics.mideast

Table 3: Data Sets Generated from Reuters-21578

Data set	Source Domain $\mathcal{D}^s$	Target Domain $\mathcal{D}^t$
Orgs vs People	Orgs.{...}, People.{...}	Orgs.{...}, People.{...}
Orgs vs Places	Orgs.{...}, Places.{...}	orgs.{...}, Places.{...}
People vs Places	People.{...}, Places.{...}	People.{...}, Places.{...}