# Second Year Project
# Cross-Domain Sentiment Classification Using Feature Expansion

**Shakir Shaker**
IT University
Copenhagen, Denmark
shsh@itu.dk

**Sofia Elena Terenziani**
IT University
Copenhagen, Denmark
sote@itu.dk

## Abstract

Binary sentiment classification deals with the task of training a binary classifier using data annotated for positive or negative sentiment. Sentiment is expressed differently in different domains and annotated corpora for every domain of interest is expensive. Consequently, it would be beneficial if a model trained on a domain, could perform comparably well with data from another domain. This paper aims at improving an initial constructed baseline to be feasible to perform cross-domain sentiment classification. To overcome the challenges of cross-domain classification the proposed method uses feature expansion. A glossary is created to align words expressing equal sentiment across domains. The glossary is used to expand feature vectors of reviews in both source and target domain. The new features are used to learn a binary classifier.

## Introduction

As we see an exponential increase in the availability and popularity of online reviews, the task of sentiment analysis has become an interesting topic in research. However, customer reviews can cover various types of products, making it difficult to gather annotated corpora for each of them. The collection of annotated data is a time and resource expensive and not scalable task as the amount of data increases. Cross-domain learning focuses on the task of adapting the knowledge learned from one or more source domains to a different target.

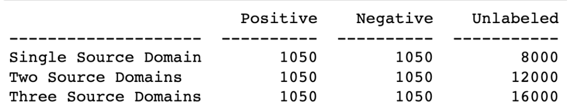
The paper is based on a constructed baseline, aimed at predicting the sentiment polarity of user-generated sentiment data in a specific domain, namely music reviews. The baseline uses CountVectorizer for text vectorization and logistic regression as classification model. Motivated by the above-mentioned need to apply knowledge across domains, we address the following research question: How can we improve our baseline sentiment classifier to be feasible to perform a cross-domain sentiment classification task? Inspired by previous work, we model the cross-domain sentiment classification task as one of feature expansion [1]. We propose a model that appends additional related features to the feature vectors that represent each review at training and test time. The main challenges of transfer learning across domains lie in the sparsity of data distributions and word semantics. In cross-domain settings, knowledge may not be directly transferred from one domain to another. The proposed model overcomes the challenges of cross-domain classification by expanding each review features with words that express equal sentiment across domains, hence reducing the feature mismatch between domains. The potential expanded features are collected from a glossary, which aligns and captures the relatedness of words used across domains.

## Related Work

Pan et al. [2] propose the SFA algorithm to reduce the gap between domains. The key intuition behind the SFA algorithm is to align domain-specific words from different domains into unified clusters and use domain-independent words as a bridge between them. The SFA algorithm constructs a bipartite graph and adapts a spectral clustering algorithm to co-align the sets of features into clusters. The clusters present the new lower-dimensional representation of all data samples and are used to learn a binary classification model [2]. Blitzer et al.[3] propose the Structural Correspondence Learning (SCL) algorithm to learn a common feature representation that is meaningful across domains. SCL aims at identifying correspondences among features from different domains by modeling their correlations with words that behave similar across domains, defined as pivot features. The key intuition behind SCL is that even when non-pivot features are completely district across domains, if they present a high correlation with the same pivot feature, then they can be treated similarly and therefore aligned to form a new feature space. A binary classifier is then learned on the new feature representation[3]. SCL and SFA are often referred to as the state-of-art domain adaptation technique. Both rely on learning pivot or domain-independent features to bridge the gap between domains and create a feature space that is common in both domains. Bollegala et al.[1] proposed a cross-domain sentiment classification method using an automatically created sentiment-sensitive thesaurus (SST). The proposed method uses labeled data from the source domain and unlabeled data from the target domain to create a thesaurus that groups words expressing the same sentiment across domains. The SST is used to expand the ultimate feature vector in order to mitigate the sparseness of data. The expanded vectors are then used to learn a binary classifier. It is the work of Bollegala et al.[1] that has inspired our research.

## Data

We define the task of cross-domain sentiment classification as the one of adapting the knowledge learned labelled data from one or more source domains to a different target domain. In addition to labeled data, we assume the availability of unlabeled data from both source and target domain. The project uses user-generated sentiment data provided by Amazon [4]. Review data are collected from four different product types (books, music, electronics, and pet supplies). Each review in each Amazon dataset is assigned a rating, ranging from 0 to 5. Reviews with ratings over 3 are labeled as positive (1) whereas those with ratings lower than 3 are labeled as negative (0). Ratings equal to 3 are removed from the dataset, obtaining binary sentiment datasets. Moreover, only reviewed texts and review summaries are considered from the provided datasets. Together they represent one review instance.



```
                      Positive    Negative    Unlabeled
--------------------  ----------  ----------  -----------
Single Source Domain        1050        1050         8000
Two Source Domains          1050        1050        12000
Three Source Domains        1050        1050        16000
```

Figure 1: Number of instances for each domain adapted

To address the research question, we select music review as the target domain for all conducted experiments. The remaining domains are used in turn as single source domains or are combined to multiple source domains. From the source labeled data, 2100 instances are selected. The instances are equally balanced according to sentiment label. Moreover 4.000 unlabelled instances from both target and source domains are selected. When source domains are combined we also limit the number of labeled instances to 2100, equally balanced according to sentiment label and domain origin. The limit enables us to perform a fair evaluation when comparing the performance of a single source domain and multiple source domains. Moreover 4.000 unlabelled instances are selected for each source domain taken into consideration. Following the standard, 420 instances from the music review are selected as the target domain for all conducted experiments.

## Methodology

Cleaning and preprocessing are conducted for all given labeled and unlabeled reviews. Firstly, removal of punctuations and stop words is performed for each instance. Next lemmatization is performed to reduce words from each review to a normalized form. N-grams i term of unigrams and bigrams are retrieve from each review using the ngrams function present in the nltk library[5].

The glossary is created using the labeled data from the source domain and unlabelled data from both target and source domain. Especially unlabeled data play an important role in the creation of the glossary as it provides accurately estimates of the distribution of words across domains. The glossary is created for both unigrams and bigrams separately. For sake of simplicity, we refer to both as n-grams for the remanding of this section.

The proposed method is built upon the Distributional Hypothesis. The hypothesis states that words are semantically related if they have many common co-occurring words[6]. The main idea is that words that occur in the same context tend to have similar meaning, implying that words are mainly characterized by the company they keep. Following the distributional hypothesis, we model each n-gram according to the n-grams they co-occur with. To leverage computational time, a new vocabulary of n-grams is created representing the most frequent n-grams in the dataset. The maximum frequency of considered elements in the new vocabulary is calculated over the total number of elements appearing in the training set. It is different between unigrams and bigrams, as individual bigrams are expected to occur much less than individual unigrams. The co-occurrence of each element of the new obtained vocabulary is represented by a term-context matrix. The matrix is created using the k-skip-n-gram approach, which defines the context as a sliding window of k+n elements. Elements that co-occurs within the context are said to have co-occurred. The value of the co-occurrence is weighted using Point-Wise Mutual Information (PMI). By quantifying the likelihood of the co-occurrence of two elements, PMI is used as a statistical measure to assess whether the co-occurrence is meaningful or not [7]. PMI computes the log probability of co-occurrence scaled by the product of single probability of occurrence.

$$\text{PPMI}(w, c) = \log \frac{\text{P}(w, c)}{\text{P}(w)\text{P}(c)}$$

PMI implies that if either one of the elements has a low probability of occurrence, but the joint probability of the two elements is high, the co-occurrence of the two elements is higher than what we would expect if they would be independent events. For the proposed model only positive pointwise mutual information (PPMI) scores are considered, as we do not want to model elements that are co-occurring less than expected by chance (less than 0)[7].

$$\text{PPMI}(w, c) = \max(0, \text{PMI}(w, c)$$

In addition to the co-occurrence features, we make use of sentiment labels where possible. For each labeled source domain review we append the label of the review to each n-gram. We use the notation $\backslash_p os$" to indicate positive sentiment elements and $\backslash_n eg$" to indicate negative sentiment. The method is proposed by Bollegala et al. [1] to encode sentiment information to form a part of the context features. By computing the co-occurrence over n-grams with appended features, we are able to create a representation of each n-gram that is characterized by the elements that they appear semantically close to and is sensitive towards the sentiment expressed in the context they appear in.

Given two n-grams and their respective feature vectors, we model their relatedness using the Cosine Similarity Measure. The Cosine Similarity Measure has been widely used in the NLP task of semantic textual similarity [8] [9]. It computes the angle between two vectors, indicating their similarity or dissimilarity in a range from -1 to 1.

$$\text{sim}(w, c) = \frac{w \cdot c}{\parallel w \parallel \parallel c \parallel}$$

The Cosine Similarity score of two n-grams is interpreted as the proportion of the features weighted according to PMI that are shared across n-grams. The relatedness measure is used to construct a glossary in which for each n-gram $w$, we list up n-gram elements $c$ that co-occur with it, PPMI$(w, c) > 0$, in descending order of relatedness values, sim$(w, c)$. The glossary is represented as a dictionary presenting each n-gram in the vocabulary as keys and the related and co-occurring elements as values.

Feature expansion is motivated yet again by the Distributional Hypothesis [6]. All keys of the glossary are interpreted as potential features and the values of each key as words that co-occurs and are semantically related to the key. If a review presents a prominent amount of n-grams that are equal to the values of a particular key, we assume that the key must be semantically related to the review. The count of equal words is normalized by the length of each review and if it passes a threshold, the key is appended to the review. While the glossary is created solely from the training instances, feature expansion is performed both at training and test time. The described relationship is computed both for the source domain training instances and for the target domain test instances. The result is new data sets both for target and source domains, where additional elements are appended to the original reviews if these are assumed to be co-occurrent and related to the review.

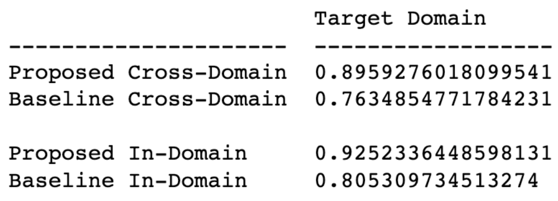
Following the research question, we implement the same classification method as performed in our baseline. Bag of words (BoW) method is used as the feature text representation for each instance. The bag of words are generated using CountVectorizer from Sklearn[10], where the n-gram range is increased to extract both unigrams and bigrams counts. The vectorization converts the collection of text reviews to a matrix of unigram and bigram counts. Logistic Regression is performed as the binary classification algorithm. It is worth mentioning that the proposed method is agnostic to the properties of the classifier and can be used for any binary classifier. The model is evaluated using the accuracy score, the set of labels predicted for a sample that exactly match the corresponding set of labels in the true sentiment class.

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

We make use of the accuracy score metric as the binary sentiment classes are balanced and equally important in all conducted experiment.

## Results

Results show that the proposed method outperforms the baseline model in each conducted experiment. Figure 2 shows the results of the performance of the proposed method against the initial baseline model. The table shows the result in the case where multiple sources domains are combined both for the proposed method and for the baseline. Moreover, the table shows the results when performing classification in-domain, the setting in which music reviews are used both as source and target domain. To study the effect of using multiple or combined source domains a classifier is trained using all possible combinations of the three source domains: pet supplies (P), books (B) and electronics (E). A total of 14 experiments have been conducted. Figure 4 and Figure 5 show the results of the baseline model and the proposed model respectively for each 14 experiment. The proposed method is built upon the construction of the glossary and its performance directly depends on it.



```
                         Target Domain
---------------------    ------------------
Proposed Cross-Domain    0.8959276018099541
Baseline Cross-Domain    0.7634854771784231

Proposed In-Domain       0.9252336448598131
Baseline In-Domain       0.805309734513274
```

Figure 2: Results of proposed and baseline model

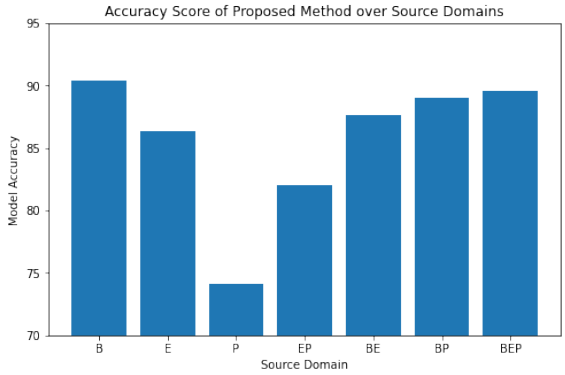


Figure 3: Results of Proposed Model

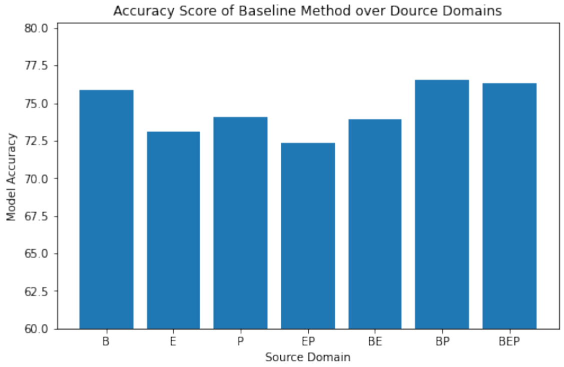


Figure 4: Results of Baseline Model

For constructing a meaningful glossary, parameter tuning is required. The results of an evaluation of the size of the glossary over the accuracy of the conducted experiment on multiple source domains is shown in Figure 5.

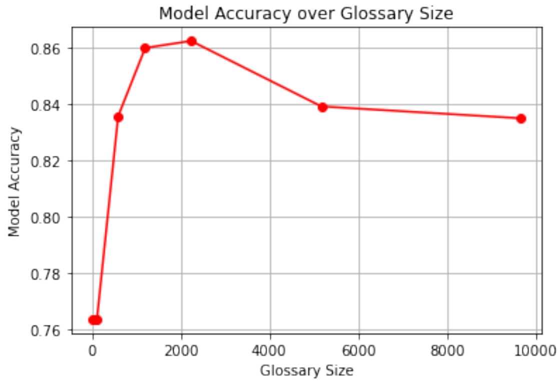


Figure 5: Evaluation of glossary size

## Discussion And Error Analysis

The results show that the creation of glossary for feature expansion is useful to perform cross-domain sentiment classification. The proposed model outperforms the baseline model with no domain adaptation in all conducted experiments, suggesting that the proposed method is able to reduce the gap between domains and learn additional features that are relevant for learning a strong classifier. However, the proposed model does not perform cross-domain classification as well as in-domain classification. Reasons for the classification error can again be assigned to the challenges we may encounter when learning across domains. Data in source and target domains are often differently distributed, making it hard for a model trained on the source domain and its underlying feature distribution to make valuable predictions on another target domain.

Selecting the correct source domains to adapt the model to a specific target domain is a challenging task. Overall, the use of multiple source domain is recommendable, as it averages the performance of single source domains. Moreover, it would decrease the amount of time spent on selecting the most feasible source domain to adapt to a given target domain. The results show that books is the single best source domain when adapting the model to the music target domain. This is true both for the model trained using the proposed method and for the baseline. The behavior is explained by the fact that in general books and music have similar aspects. On the other hand, pet supplies acts as the worst source domain when adapting the model to the music target domain. The model trained on a single pet supplies source domain is the only one that does not seem to have benefited from the proposed model. The trend is also seen when combining source domains with the pet supplies data set. The magnitude of the drop in performance is related to the degree of similarity between domains. An assumption could be that the pet supplies review dataset does not present enough diversity of evaluative expression, causing the model to fail at carrying useful information to the target domain. The trend is although not true for the baseline model, suggesting that the features extracted from the pet supplies dataset are not as useful as the one collected from the other source domains.

The performance of the proposed method directly depends on the glossary used for feature expansion. Experiments on the effect of the number of keys selected by the model show that initially, when the number of keys in the glossary is increased, the classification accuracy increases. This is seen as the result of feature expansion enabling the model to reduce the mismatch between source and target domain features. However, when the size of the keys is further increased, the model's performance drops and saturates. An explanation is seen as the size of the glossary increases, we would also increase the diversity of candidates for the feature expansion. Diversity is also expected to be introduced when we increase the number of neighboring values for each key in the glossary.

Feature expansion uses a threshold when deciding whether a feature should be appended to the review or not. This is another example of parameters that needs to be tuned in order to obtain a meaningful glossary. The threshold is highly dependent on the size of the training data and especially on the size of the glossary and can cause great variation in the performance of the model.

## Concluding Remarks And Future Work

The project is based on a constructed baseline, aimed at predicting the binary sentiment orientation of music reviews. We propose an approach to further build upon the baseline, enabling it to perform a robust cross-domain sentiment analysis. The proposed method firstly creates a glossary to identify co-occurrence, similarity and distributional relatedness among words. The glossary is used to expand review features, which are further used to train a binary classifier. Feature expansion enables the model to bridge the gap between domains by incorporating additional related features to feature vectors that represent source and target domain reviews. By appending words that co-occur and are related according to statistical metrics, the model is able to append features that originate from different source domains, bridging the gap across domains. The performance of the proposed model depends on the glossary used for feature expansion. As seen in the results, the model performs poorly with domains that are semantically distant from each other. For further work we would like to research how to improve the model so it performs well on domains considered distant to each other. Moreover, the glossary needs a lot of parameter tuning steps. For further work we would like to investigate how we could mitigate this, by applying evaluation of parameters during the creation of the glossary.

## References

[1] Danushka Bollegala, David Weir, and John Carroll. Cross-domain sentiment classification using a sentiment sensitive thesaurus. *IEEE Transactions on Knowledge and Data Engineering*, 25(8):1719–1731, 2013.

[2] Sinno Jialin Pan, Xiaochuan Ni, Jian-Tao Sun, Qiang Yang, and Zheng Chen. Cross-domain sentiment classification via spectral feature alignment. In *WWW '10*, 2010.

[3] John Blitzer, Ryan McDonald, and Fernando Pereira. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 120–128, Sydney, Australia, July 2006. Association for Computational Linguistics.

[4] Jianmo Ni, Jiacheng Li, and Julian McAuley. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 188–197, Hong Kong, China, November 2019. Association for Computational Linguistics.

[5] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.", 2009.

[6] Magnus Sahlgren. The distributional hypothesis. *The Italian Journal of Linguistics*, 20:33–54, 2008.

[7] Valentina Alto. Understanding pointwise mutual information in nlp, January 2020. [Online; posted 31-January-2020].

[8] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.

[9] Richmond Alake. Understanding cosine similarity and its applications, September 2020. [Online; posted 15-September-2020].

[10] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

## Github Repository

https://github.itu.dk/sote/SecondYearProject15

## Results of phase 1 and phase 2

We briefly describe how the baseline is constructed. Preprocessing and cleaning is performed firstly removing all rows that presented missing values for the review and summary columns. Stropwords and punctuation are then removed from each review. Each review is split into tokens using the word tokenizer function available in the nltk library. We use CountVectorizer for text vectorization, where the n-gram range is increased to extract both unigrams and bigrams counts. The vectorization converts the collection of text reviews to a matrix of unigram and bigram counts. We chose Logistic Regression as the binary classification model. Model gave us 0.9125 accuracy for predictions on dev dataset. The predictions for test cases and hard cases are uploaded to CodaLab for Group 15. The accuracy of the predictions on the test cases were shown in CodaLab to be close to 0.90, while for the hardcases the accuracy dropped by almost 20