International Research Journal of Engineering and Management Studies (IRJEMS)
Volume: 03 Issue: 05 | May -2019 ISSN: 2395-0126

CROSS DOMAIN TEXT CLASSIFICATION USING LABELLED LDA METHOD

Lavanya R[1], Mrs.K.Krishnakumari[2]
[1]*Student, Department of Computer Science and Engineering*
A.V.C. College of Engineering,
Mannampandal, Mayiladuthurai – 609305
lavyyraju@gmail.com
[2] *Associate Professor, Department of Computer Science and Engineering*
A.V.C. College of Engineering,
Mannampandal, Mayiladuthurai – 609305
krishna.41999@gmail.com

Abstract - *Cross domain text classification aims at building a classifier for a target domain which leverages data from both source and target domain. To address the problem of restricting model's learning ability and impairing model's performance on classification tasks when the semantic distributions of different domains are very different, we propose a novel group alignment which aligns the semantics at the group level based on topic. In addition to help the model learn better semantic groups and semantics within these groups, we also propose an effective method of partial supervision for model's learning in source domain. Then, we embed the group alignment and a partial supervision into a cross-domain topic model, and propose a Cross Domain Labeled LDA (CDL-LDA) which improves the accuracy of cross domain adaptation problems.*

Keywords: text classification, group alignment, topic modelling, supervision.

1.INTRODUCTION

The widely used natural language processing task in different business problem is text Classification. The goal of text classification is classify the text documents into one or more defined categories automatically and to increase discoverability of information and availability for all the knowledge discovered or actionable to support the strategic decision making. Text classification is an effective supervised machine learning method and widely used for the purpose of classifying the sentences or text documents into one or more defined categories. It is widely used for natural language processing task which is playing an important role in spam filtering, sentiment analysis, categorization of news articles and also many other business related issues.

Text classification systems have been adopted by the growing number of organizations to manage the ever growing inflow of unstructured information effectively.

Some examples of text classification are:

- Understanding audience sentiment from social media,
- Detection of spam and non-spam emails.

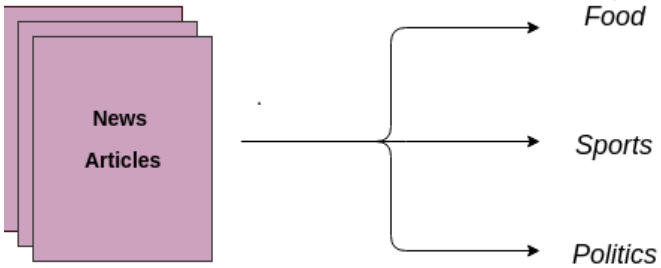

Figure 1 text classification

1.2 OVERVIEW

Cross-domain Labeled LDA (CDL-LDA) is a cross-domain topic model which divides topics into common topics an-d specific topics to model shared semantics across domains and domain dependent semantics respectively. To perform the task of text classification, it is necessary to align specific topics across domains. However, different from exact alignment, which directly performs topic alignment at topic level, we propose a novel group alignment which performs topic alignment at topic group level. Traditional machine learning algorithms often fail to obtain satisfactory performance when the training and test data are drawn from different but related data distributions. See the task example as follows.

In addition to group alignment, we also propose a partial supervision to explicitly incorporate ground truth topic group information in the source domain, which can help the model learn better topic features for classification.

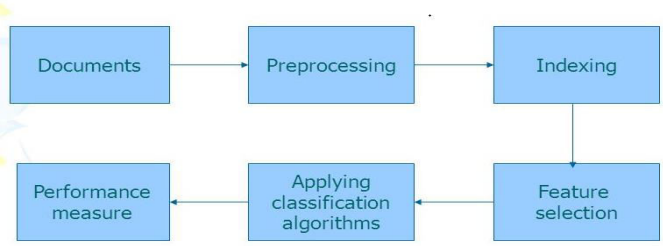

Figure 2 Steps of text classification

2. RELATED WORK

Pragya Tripathi et al Text mining is basis for the analysis of data in natural language text. It is used to process unstructured (textual) information, extract meaningful numeric indices from the text. Tokenization that splits the text of a document into a sequence of tokens. Transform Cases that transforms all characters in a document to either lower case or upper case, respectively. **Lianghao Li et al** propose a novel approach named Topic Correlation Analysis (TCA) for cross-domain text classification. TCA extracts both the shared and the domain-specific latent features as a bridge to transfer the text classification knowledge between domains. **Desheng Dash Wu et al** Important questions include: Is there any obvious correlation between investor sentiment and stock market performance? How can investor sentiment predict future stock price? In order to answer these questions, we use sentiment related index. Research on sentiment index selection has identified both direct and indirect sentiment indices. However, a direct sentiment index based on questionnaires and an indirect sentiment index relies on related stock market data would be inaccurate. **Thomas L. Griffiths et al** Here, statistical methods are used for automatically extracting a representation of documents that provides a first order approximation to the kind of knowledge available for the domain experts. Then the statistical model was highly reducing the complex process of producing a scientific paper into small number of simple probabilistic steps which is more effective and then it specifies that the probability distribution. **Wenyuan Dai, et al** In this paper, main focus is on the problem of classifying documents across different domains. We now describe the co-clustering based classification algorithm for classifying the out-of-domain data, which minimizes the objective function. The objective function is a multi-part function which is hard to be optimized. **Shokoufeh Salem Minab et al** In this research, selecting correct keywords can be caused appropriate classification and presenting the results show importance of preprocessing to analysis of emotions. They have used two steps include deformation and filtering. **Peter Pretten hofer et al** Standard text classification distinguishes be- tween labeled (training) documents and unlabeled (test) documents. Cross-language text classification poses an extra constraint in that training documents and test documents are written in different languages. **Yang Bao et al ,** The goal of our model is to induce a new latent feature space so that we can achieve better performance for cross-domain text classification. On one hand, it is necessary to model documents in each domain as a separate collection. On the other hand, it is crucial to consider the duality between the marginal distribution of examples and the conditional distribution of class labels given examples. **Sankar.H et al** The SA is different from traditional topic based classification techniques, in topic based classification can be performed by identifying the keyword present in the document, while the sentiment analysis need much precise information. Therefore, the sentiment analysis requires much understanding than traditional document classification technique. **David M. Blei et al** Here**,** Latent Dirichlet allocation (LDA) was proposed, in that a generative probabilistic model for collections of more discrete data such as text corpora. LDA is a method of three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics in the document.

2.1 DISADVANTAGES OF EXISTING SYSTEM

- Restricts the model's learning ability.
- Impairing model's performance on classification tasks when the semantic distributions of the different domains can be very different.
- Mainly, Specific topics of the target domain can be decomposed by the specific topics of the source domain.

3. PROPOSED SYSTEM

Cross-Domain Labeled LDA (CDL-LDA) model for cross-domain text classification which is equipped with a novel group alignment which performs topic alignment at topic group level. To be more specific, firstly both the common topics and specific topics are partitioned into different groups. Then within each domain, common topics and specific topics of the same group are aligned. Finally, across domains, specific topics of the same topic group are aligned through common topics in this topic group. In addition to group alignment, we also propose a partial supervision to explicitly incorporate ground truth topic group information in the source domain, which can help the model learn better topic features for classification. To do so, instead of sampling topic group labels for words in source domain, the model directly assign ground truth topic group labels of the words at training time.

3.1 ADVANTAGE OF PROPOSED SYSTEM

- The groups are guaranteed to exist in both of the source and the target domains, thus aligning topics by groups are always feasible.
- The numbers of topics within different groups are allowed to be different, and thus the model will have more flexibility for modeling topics in different domains.

4. SYSTEM DESIGN

In this work, first taking the dataset of ABS newsgroup of different domains, then classifying the text by the method of labelled LDA which comprises of both group alignment and partial supervision.

To show the effectiveness it is compared with the unsupervised CDL-LDA. Then, in Different numbers of specific topic it provides the necessary steps to better model the different semantics of different domains and improve the model's representation flexibility, CDL-LDA allows the numbers of topics to be different for different domains. And in the next step parameter analysis has to be done with the two major methods Hyper parameter and number of topics to evaluate the accuracy. Then in topic detection we qualitatively evaluate group alignment and the proposed partial supervision adopted by CDL-LDA through topic detection experiment.

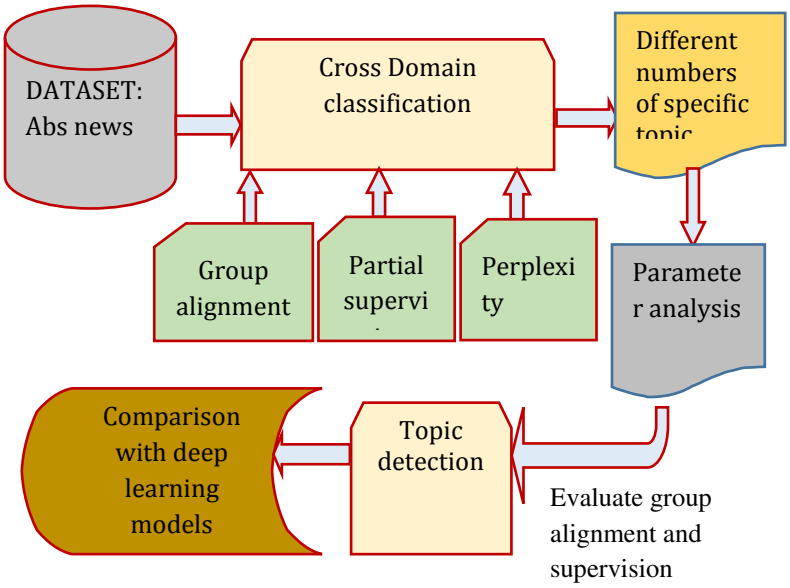

Figure 3 System architecture

Then for comparison with deep learning approach we select two sets of deep learning models: 1) Auto Encoder based models: marginalized Stacked Denoising Auto Encoder (mSDA)1-norm Stacked Robust Auto Encoder 2) Domain Adversarial Neural Network based models: Domain-Adversarial Neural Network (DANN) and Adversarial Representation Learning for Domain Adaptation (ARDA).

4.1 MODULES

In this project there are five modules are used to classify the sentiment in the review dataset. The modules are,

1. Dataset collection
2. Preprocessing
3. Cross domain classification
4. Perplexity
5. Different number of specific topics
6. Parameter analysis
7. Topic detection
8. Comparing with deep learning models

SAMPLE DATASET

Publish date	Headline text
20030219	aba decides against community broadcasting license
20030219	act fire witnesses must be aware of defamation
20030219	a g calls for infrastructure protection summit

Table 1 Example data

4.1.1 DATASET COLLECTION

The data is taken from across different ABC digital platforms during the month of Feb 2003. These platforms include all the data from ABC News desktop and mobile websites and the ABC app (both iOS and Android versions).

4.1.2 PREPROCESSING

Tokenization - Segregation of the text into its individual constituent words.

Stop words - Throw away any words that occur too frequently as its frequency of occurrence will not be useful in helping detecting relevant texts.

Stemming - combine variants of words into a single parent word that still conveys the same meaning

Vectorization - Converting text into vector format. One of the simplest method is the famous bag-of-words approach, where you create a matrix .

4.2.2 CROSS DOMAIN CLASSIFICATION

We conduct two sets of cross domain classification experiments: binary classification and 4-class classification. In the binary classification tasks, we compare CDL-LDA with several state-of-the-art models: SFA, TPLSA, CDPLSA, TCA and PSCCLDA. In the 4-class classification tasks, we compare CDL-LDA with state-of-the-art cross-collection topic models TCA and PSCCLDA.

GROUP ALIGNMENT

First, common topics and specific topics are firstly divided into different groups. Then within each domain, common topics and specific topics of the same group are aligned. Finally, across domains, specific topics of the same topic group are aligned through common topics in this topic group.

PARTIAL SUPERVISION

Unsupervised models usually ignore the valuable information provided by the labels of training data. For classification tasks, supervision can help model to learn better features for classification. To show the effectiveness of the proposed partial supervision method, we compare CDL-LDA with CDL-LDA^{un}.

4.2.3 PERPLEXITY

Perplexity is a popular evaluation metric for topic models, and a lower perplexity indicates a better representation or generalization ability of the model. Here,

we adopt perplexity to evaluate model's generalization ability on the target domain. Perplexity is calculated through the following equation:

$$P(D^{tgt}|D^{src}) = exp(-\frac{\sum_{d=1}^{|D^{tgt}|} \log p(D_d^{tgt}|D^{src})}{\sum_{d=1}^{|D^{src}|} N_d})$$

where D^{src} and D^{tgt} are documents from the source domain and the target domain respectively; $|D^{src}|$ and $|D^{tgt}|$ are the number of documents in the source and the target domains; N_d denotes the number of words in document d.

4.2.4 DIFFERENT NUMBER OF SPECIFIC TOPICS

To better model the different semantics of different domains and improve the model's representation flexibility, CDL-LDA allows the numbers of topics to be different for different domains. By adopting different number of topic in different domains, CDL-LDA can better model different semantics of both source and target domains, and achieve better classification performances.

4.2.5 PARAMETER ANALYSIS

This section presents experiments aimed at testing the influences of different parameters in CDL-LDA. We have four hyper-parameters in CDL-LDA, including α, β, γ and η, and two parameters about the number of topics: total number of topics TC + TS and the ratio of common topics TC. We evaluate the influence of these parameters on the 20Newsgroups dataset.

4.2.6 TOPIC DETECTION

In this section, it qualitatively evaluate group alignment and the proposed partial supervision adopted by CDL-LDA through topic detection experiment. Such behavior characterizes the group alignment : it only align topics at topic group level instead of topic level.

4.2.8 COMPARING WITH DEPP LEARNING MODELS

First it select two sets of deep learning models: 1) Auto Encoder based models: marginalized Stacked Denoising Auto Encoder (mSDA) and $l_{2,1}$- norm Stacked Robust Auto Encoder . 2) Domain Adversarial Neural Network based models: Domain Adversarial Neural Network (DANN) and Adversarial Representation Learning for Domain Adaptation (ARDA).

It can observe that the proposed CDL- LDA not only outperforms state-of-the-art cross-collection topic models with exact alignment, but also the sate-of-the-art deep learning models which also adopt exact alignment, which demonstrates the effectiveness of the proposed group alignment and the proposed partial supervision.

5. RESULTS AND DISCUSSION

CDL-LDA outperforms all of these state-of-the-art methods on the rest of tasks and improves the classification accuracies by [1.7%, 5.7%]. On average, CDL-LDA improves the accuracy from 88.0% (PSCCLDA) to 90.5%. In the 4-class classification tasks, we compare CDL-LDA with state-of-the-art cross-collection topic models TCA [19] and PSCCLDA. We can observe improvements of [7.2%, 14.6%] on different tasks, and an average improvement of 11.4%. These improvements indicate the effectiveness of proposed group alignment and partial supervision.

In group alignment, CDL-LDA^{un} outperforms CCLDA on all of the tasks, and improves classification accuracies by [2.2%, 16.9%] for binary classification tasks and [13.1%,25.1%] for 4-class classification tasks. Besides, CDL-LDA^{un} improves averaged accuracies from 75.0% to 82.2% and from 51.0% to 71.1% on binary and 4-class classification tasks respectively.

In partial supervision, we can observe significant increases of classification accuracies made by the proposed partial supervision. We can also observe improvements of [2.2%, 16.3%] on different binary classification tasks, and improvements of [13.1%,25.1%] on 4-class classification tasks. On average, CDL-LDA improves classification accuracies over CDL-LDA^{un} by 8.3% and 18.0% on binary and 4-class classification dataset respectively.

6. CONCLUSION

Here, we proposed Cross-Domain Labeled LDA (CDL-LDA) for cross-domain text classification, along with a group alignment and a partial supervision. Then the quantitative experiments show that both the group alignment and the partial supervision can help model learn better features for both classification and generalization and qualitative experiment shows that the proposed model is able to not only detect meaningful topics, but also successfully align topics at topic group level.

7. FUTURE ENHANCEMENT

In our future work, we plan to explore alternate methodologies to leverage and organize the common language substrate of the given domains. We also plan to extend our approach to perform cross-language text classification, an interesting problem with difficult challenges.

REFERENCES

[1] Li, L., Jin, X., & Long, M. (2012, July). Topic correlation analysis for cross-domain text classification. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*.

[2] Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences*, *101*(suppl 1), 5228-5235

[3] Bao, Y., Collier, N., & Datta, A. (2013, October). A partially supervised cross-collection topic model for cross-domain text classification. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management* (pp. 239-248). ACM.

[4] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993-1022.

[5] Dai, W., Xue, G. R., Yang, Q., & Yu, Y. (2007, August). Co-clustering based classification for out-of-domain documents. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 210-219). ACM.
[6] Prettenhofer, P., & Stein, B. (2010). Cross-language text classification using structural correspondence learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics* (pp. 1118-1127).
[7] Sankar, H., & Subramaniyaswamy, V. (2017, December). Investigating sentiment analysis using machine learning approach. In *2017 International Conference on Intelligent Sustainable Systems (ICISS)* (pp. 87-92). IEEE.
[8] Minab, S. S., Jalali, M., & Moattar, M. H. (2015, November). Online analysis of sentiment on Twitter. In *2015 International Congress on Technology, Communication and Knowledge (ICTCK)* (pp. 359-365). IEEE.
[9] Tripathi, P., Vishwakarma, S. K., & Lala, A. (2015, December). Sentiment analysis of english tweets using rapid miner. In *2015 International Conference on Computational Intelligence and Communication Networks (CICN)* (pp. 668-672). IEEE.
[10] Li, N., & Wu, D. D. (2010). Using text mining and sentiment analysis for online forums hotspot detection and forecast. *Decision support systems*, *48*(2), 354-368.