International Journal of Engineering and Techniques - Volume 3 Issue 1, Jan – Feb 2017

RESEARCH ARTICLE OPEN ACCESS

Emotion Detection from Text Based Document and classification of Cross-Domain Sentiment

Mr.S.Parthiban[1], Mrs.T.Priyadarsini[2]

[1]Assistant Professor, Department of Electronics & Communication Engineering,ISTE Member,

[2]Assistant Professor, Department of Electronics & Communication Engineering,

V.R.S. COLLEGE OF ENGINEERING & TECHNOLOGY, Arasur – 607107, Villupuram District, Tamilnadu.

ISTE Member, IAENG Member, IRED Associate Member

Abstract:

Sentiment analysis focuses on analyzing web documents, especially user-generated content such as product reviews, to identify opinionated documents, sentences and opinion holders. Most of the time classifiers trained in one domain do not perform well in another domain. The existing approaches do not detect sentiment and topics simultaneously. Sentiments may differ with topics. Our proposed model called Joint Sentiment Topic (JST) model to detect sentiments and topics simultaneously from text. This model is based on Gibbs sampling algorithm. Besides, unlike supervised approaches to opinion mining which often fail to produce good performance when shifting to other domains, the semi-supervised nature of JST makes it highly portable to other domains. JST model performs better when compared to existing supervised approaches.

Keywords— **opinion mining, domain adaptation, semi-supervised.**

1. INTRODUCTION

Online users are getting increased nowadays because of the fast growth in technology and their willingness to engage in social interactions. Users express their opinion about the product or goods they consume in the review site or that particular product site. This leads to the development of emotion oriented services. It is useful for both the consumers as well as for the producers to know what general public think about a particular product or service.

Given a review based document, sentiment classification aims to classify the document with respect to the sentiment expressed by the author of the review. Semantic classifier classifies the review document based on polarity as positive, negative or neutral. It provides a new aspect for document categorization, product reputation mining, and customer opinion extraction/summarization and sentiment classification. The producer also gains information about the view and survival of the product in the market and the considerable changes needed to sustain in market and to satisfy customers.

Sentiment classification based on Machine Learning requires a large amount of human annotation. One of the main challenges for sentiment classification is the domain adaptation problem.

Sentiment is expressed differently in different domains. To annotate corpora for every possible domain is a difficult task. Applying a sentiment classifier trained using labelled data for a particular domain to classify user reviews on a different domain often results in poor performance because words that occur in the train (source) domain might not appear in the test (target) domain.

2. RELATED WORK:

Sentiment analysis the task of classifying the polarity of a given text at the document, whether the expressed opinion in a document is positive, negative, or neutral. In this section, we investigate the work which deals with computational treatments of sentiment using different machine learning techniques, with a focus on document-level sentiment classification.

2.1.1 Supervised Approaches

Most supervised sentiment classification approaches use standard machine learning techniques such as support vector machines (SVMs) and Naive Bayes (NB) classifiers. These approaches are corpus-based, in which a domain-specific classifier is trained with labelled training data.

The work on document-level sentiment classification employed machine learning techniques including SVMs, NB and Maximum Entropy to determine the polarity of the sentiment expressed in the review. They

achieved the best classification accuracy with SVMs using binary features coding whether a unigram was present or not. In subsequent work further improved sentiment classification accuracy on the dataset using a cascaded approach. Instead of training a classifier on the original feature space, they first filtered out the objective sentences from the dataset using a global min-cut inference algorithm, and then used the remaining subjective sentences as input for sentiment classifier training. The classification improvement achieved by the cascaded approach suggests that the subjective sentences contain features which are more discriminative and informative than the full dataset for sentiment classification. The review dataset used in has later on become a benchmark for many sentiment classification studies. By training a SVM classifier on the combination of different types of appraisal group features and bag-of-word features, they achieved the best accuracy on the movie review dataset.

Fully supervised structured model for joint sentence and document-level sentiment classification based on sequence classification techniques using constrained inference. The joint model leverages both document-level and sentence-level label information, and allows classification decisions from one level (e.g. the document level) to influence decisions at another level (e.g. the sentence level). It was reported that the joint model significantly outperformed both the document- and sentence-classifier that predict a single level label only.

2.2 Semi-supervised approaches

2.2.1 Domain Adaptation

There has been a significant amount of work on domain adaptation for sentiment classification explored various strategies for training SVM classifiers for the target domain lacking sufficient labelled data, where training data is obtained from other domains with rich labelled examples. These strategies include (1) training on a mixture of all the available labelled data (also used as baseline); (2) training on all the available labelled data, but limiting the set of features to those observed in the target domains; (3) using ensembles of classifiers from domains with available labelled data; and (4) combining small amounts of unlabelled data with large amounts of unlabelled data in the target domain for classifier training with an algorithm. It was found that the approach provided the best classification accuracy of the four strategies because it can take advantage of unlabelled data in the target domain for training.

An approach for leveraged data from both source and target domain for sentiment adaptation, where the target domain data are completely unlabelled. The main idea of their approach is to use classifiers trained on source domains to label some informative documents in the target domain. Those informative documents, picked up by a relative similarity ranking (RSR) method, were then used to retrain a centroid classifier for the target domain sentiment classification. This approach outperformed the transductive SVM baseline classifier in most of the evaluation tasks, obtaining an average of more accuracy.

\

2.2.2 Cross-lingual Sentiment Classification

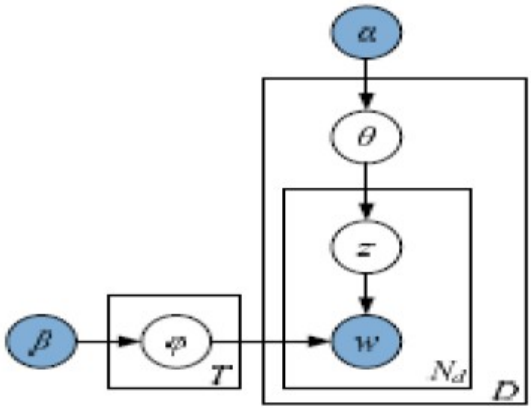
In contrast to monolingual sentiment adaptation which addresses the domain mismatch issue for sentiment classification(e.g. Book vs. Electronics reviews),cross-lingual sentiment classification focuses on the mismatch arises from language differences, that is to use labelled data from a source language to build a sentiment classifier for a different target language. Semi-supervised techniques have also been widely applied to the task of cross-lingual sentiment classification, owning to the fact that some languages (typically English) have much richer sentiment resources (e.g., labelled corpus and lexicon) than others.

2.3 Weakly-supervised Approaches

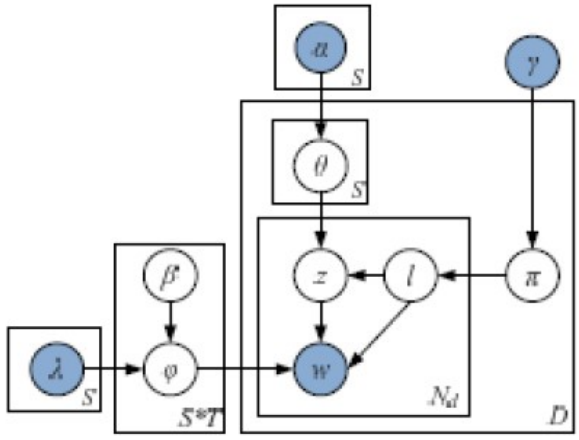
Semi-supervised approaches can be categorized as corpus-based methods as they use a labelled or unlabelled data to train sentiment classifiers. Given the difficulties of supervised and semi-supervised sentiment analysis, it is conceivable that unsupervised or weakly-supervised approaches to sentiment classification are even more challenging. Nevertheless, solutions to unsupervised or weakly-supervised sentiment classification are of practical significance owing to its domain independent nature.

3. JOINT SENTIMENT TOPIC (JST) MODEL

Document-level sentiment classification for general domains in conjunction with topic detection and topic sentiment analysis, based on the proposed weakly-supervised joint sentiment-topic. This model extends the topic model Latent Dirichlet allocation (LDA) by constructing an additional sentiment layer, assuming that topics are generated dependent on sentiment distributions and words are generated conditioned on the sentiment-topic pairs.



a) LDA Model



***b)* JST Model**

The existing framework of LDA has three hierarchical layers, where topics are associated with documents, and words are associated with topics. In order to model document sentiments, we propose a joint sentiment-topic model by adding an additional sentiment layer between the document and the topic layers. Hence, JST is effectively a four-layer model, where sentiment labels are associated with documents, under which topics are associated with sentiment labels and words are associated with both sentiment labels and topics.

Assume that we have a corpus with a collection of D documents denoted by C = {d1, d2,..., dD}; each document in the corpus is a sequence of Nd words denoted by d = (w1, w2, ..., wNd), and each word in the document is an item from a vocabulary index with V distinct terms denoted by {1, 2, ..., V }. Also, let S be the number of distinct sentiment labels, and T be the total number of topics. The generative process in JST which corresponds to the graphical model shown in figure (b) is as follows:

• For each sentiment label *l {1,...,S}*

• *For each topic j{1,....,T}, draw*

• For each document d, choose a distribution πd ~Dir (γ).

• For each sentiment label l under document d, choose a distribution θd, l ~ Dir (α).

• For each word wi in document d

– choose a sentiment label li ~ πd,

– choose a topic zi ~ (θd,li),

– choose a word wi from lazy, a distribution over words conditioned on topic zi and sentiment label li.

The hyper parameters α and β in JST can be treated as the prior observation counts for the number of times topic j associated with sentiment label l is sampled from a document and the number of times words sampled from topic j are associated with sentiment label l, respectively, before having observed any actual words. Similarly, the hyperparameter γ can be interpreted as the prior observation counts for the number of times sentiment label sampled from a document before any word from the corpus is observed. In our implementation, we used asymmetric prior α and symmetric prior β and γ. In addition, there are three sets of latent variables that we need to infer in JST, i.e., the per-document sentiment distribution π, the per-document sentiment label specific topic distribution θ, and the percorpus joint sentiment-topic word distribution '.We will see later in the paper that the per-document sentiment distribution π plays an important role in determining the document sentiment polarity.

4. EXPERIMENTAL SETUP

In this section, we present the experimental setup of document polarity classification and topic extraction based on the movie review dataset. This dataset consists of two categories of free format movie review texts, with their overall sentiment polarity labeled either positive or negative. However, one should note that we do not use any of the polarity label information of the dataset in our experiments but only for evaluating the performance of the JST model, as our model is fully unsupervised.

4.1 Pre-processing

Pre-processing was performed on the movie review data before the subsequent experiments. Firstly, punctuation, numbers and other non-alphabet characters were removed. Secondly, for the purpose of reducing the vocabulary size and addressing the issue of data sparseness, stemming was performed using the Porter's stemmer algorithm. Stop words

were also removed based on a stop word list. After pre-processing, the corpus contains 2000 documents and 627,317 words with 25,166 distinct terms.

4.2 Defining Model Priors

The sentiment classification problem is somehow more Challenging than the traditional topic-based classification, since sentiment can be expressed in a more subtle manner while topics can be identified more easily according to the co-occurrence of keywords. One of the directions for improving the sentiment detection accuracy is to incorporate prior information or subjectivity lexicon (i.e., words bearing positive or negative polarity), which can be obtained in many different ways. Some approach annotates polarity to words based on manually constructed Appraisal Groups.Other approach generates subjectivity lexicons in a semi-automatic manner.

More recently, Kaji and Kitsuregawa proposed a method which can build polarity-tagged corpus from HTML documents fully automatically. While subjectivity lexicon generation is beyond the scope of this paper, here in our experiments, we investigated incorporating prior information obtained in four different ways into the JST and the tying-JST model, and explored how the prior information can improve the sentiment classification accuracy.

4.2.1 Paradigm word list

The paradigm word list consists of a set of positive and negative words, e.g. excellent and rubbish. These lexicon words can be simply treated as the paradigms for defining the positive and negative semantic orientation, rather than for the purpose of training the algorithm .The majority of the words were derived from the word lists used by Pang et al for their baseline result tests, with punctuation like '?' and '!' removed. However, we did notice the difference that the movie review data used by Pang et al. it is an older version with only 700 positive and 700 negative movie reviews, compared to the newer version we used that contains 1000 positive and 1000 negative documents. Hence, we added some additional paradigm words to the original list by re-examining a small portion of the corpus based on a very preliminary check of word frequency counts. Finally, the resulting paradigm word list contains 21 positive and 21 negative paradigm words respectively.

4.2.2 Mutual information (MI)

In statistical language modelling, mutual information is a criterion widely used for calculating the semantic association between words. Here we use mutual information to select the words that have strong association with positive or negative sentiment classes. The top 20 words within each individual sentiment class were selected based on their MI scores and incorporated as prior information for our models.

4.2.3 Full subjectivity lexicon

We also explored using the publicly available subjectivity word list with established polarities such as the MPQA subjectivity lexicon3, which consists of 2718 positive and 4911 negative words4. By matching the words in the MPQA subjectivity lexicon with the vocabulary (with 25,166 distinct terms) of the movie review dataset, we finally obtained a subset of 1335 positive, 2214 negative words.

4.2.3 Filtered subjectivity lexicon

The filtered subjectivity lexicon was obtained by removing from the full subjectivity lexicon the words occurred less than 50 times in the movie review dataset. The words whose polarity changed after stemming were also removed. Finally ,the filtered subjectivity lexicon contains 374 positive and675 negative words. Although one may argue that the paradigm word list and the MI extracted words seem requiring certain supervision information from the corpus itself, the subjectivity lexicon used here is fully domain-independent and does not bear any supervision information specifically to the movie review dataset. In fact, the JST model with the filtered subjectivity lexicon achieved better performance than the ones using the prior information obtained from paradigm word list or MI extracted words as can be seen later. While it is well-known that sentiment classifiers trained on one domain often fail to produce satisfactory results in another domain, we speculate that the unsupervised nature of our JST model makes it highly portable to other domains.

4.3 Incorporating Model Priors

We modified Phan's Gibbs LDAþþ package3 for the Implementation of JST and Reverse-JST. Compared to the original LDA model, besides adding a sentiment label generation layer θ, we also added an additional dependency link of on the matrix of size SV, which we used to encode word prior sentiment information into the JST models. The matrix λ can be considered as a transformation matrix which modifies the Dirichlet priors of size S,so that the word prior sentiment polarity can be captured. The complete procedure of incorporating prior knowledge into the JST model is as follows: first λ, is initialized with all the elements taking a value of 1. Then, for each term ω{ 1; . . . V} in the corpus

vocabulary and for each sentiment label{1 . . . ;S}, if w is found in the sentiment lexicon, the element λlω is updated as follows.

$$\lambda_{l\omega} = \begin{cases} 1, \text{if } S(\omega) = l \\ 0, \text{otherwise} \end{cases}$$

5. CONCLUSION AND FUTURE WORK

General domain sentiment classification, by incorporating a small amount of domain independent prior knowledge, the JST model achieved either better or comparable performance compared to existing semi-supervised approaches despite using no labelled documents, which demonstrates the flexibility of JST in the sentiment classification task. Classification performance of JST is very close to the best performance of Machine learning but save a lot of annotation work. Topic information indeed helps in sentiment classification as the JST model with the mixture of topics consistently outperforms a simple LDA model ignoring the mixture of topics.

In future, we plan to generalize the proposed method to solve other types of domain adaptation tasks. The JST model only incorporates prior knowledge from sentiment lexicons for model learning. It is also possible to incorporate other types of prior information, such as some known topical knowledge of product reviews, for discovering more salient topics about product features and aspects. Another possibility is to develop a semi-supervised version of JST, with some supervised information being incorporated into the model parameter estimation procedure, such as use of the sentiment labels of reviews derived automatically from the ratings provided by users, to control the Dirichlet priors of the sentiment distributions.

REFERENCES

[1] Chenghua Lin, Yulan He, Richard Everson," Weakly Supervised Joint Sentiment-Topic Detection from Text", IEEE Transactions on Knowledge and Data Engineering,Vol 24,no 6, pp.1134-1145,2012.

[2] A. Abbasi, H. Chen, and A. Salem."Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification" in web forums. ACM Trans. Inf. Syst., 26(3):1–34, 2008.

[3] Shenghua Bao, Shengliang Xu, Li Zhang, Rong Yan, Zhong Su, Dingyi Han, and Yong Yu," Mining Social Emotions from Affective Text," IEEE Transaction on Knowledge and Data Engineering, Vol 24,no 9, pp.1658-1670,2012.

[4] T. Li, Y. Zhang and V. Sindhwani, "A Non-Negative Matrix Tri-Factorization Approach to Sentiment Classification with Lexical Prior Knowledge," Proc. Joint Conf. 47th Ann. Meeting of the ACL and the Fourth Int'l Joint Conf. Natural Language Processing of the AFNLP, pp. 244-252, 2009.

[5] Andreevskaia and S. BerglerOvercoming , "Domain Dependence in Sentiment Tagging". In Proceedings of the Association for Computational Linguistics and the Human Language Technology Conference (ACL-HLT), pages 290–298, 2008.

[6] Helmut Prendinger, Alena Neviarouskaya and Mitsuru Ishizuka"SentiFul: A Lexicon for Sentiment Analysis", Proc. Fourth Int'l Workshop Semantic Evaluations (SemEval'07), pp. 70-74, 2007.

[7] Argamon, K. Bloom, A. Esuli, and F. Sebastiani."Automatically Determining Attitude Type for Multi-Domain Sentiment Analysis". Human Language Technology. Challenges of the Information Society, pages 218–231, 2009.

[8] P.C. Chang, M. Galley, and C. Manning, "Optimizing Chinese Word Segmentation for Machine Translation Performance," Proc.Assoc.for Computational Linguistics (ACL) Third Workshop Statistical Machine Translation, 2008.

[9] R. Tokuhisa, K. Inui, and Y. Matsumo "Emotion Classification Using Massive Examples Extracted From The Web," .Proc. 16th Int'l World Wide Web Conf. (WWW '07), 2007.

[10] W.H. Lin, E. Xing,and A. Hauptmann,"A Joint Topic and Perspective Model for Ideological Discourse, " Proc.European Conf. Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD '08), pp. 17-32, 2008.

[11] G. Mishne, K. Balog, M. de Rijke, and B. Ernsting, "Moodviews: Tracking and Searching Mood-Annotated Blog Posts," Proc. Int'l AAAI Conf. Weblogs and Social Media (ICWSM '07), 2007.

[12] A.M. Popescu and O. Etzioni, "Extracting Product Features and Opinions from Reviews," Proc. Joint Conf. Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP '05), pp. 339-346, 2005.