IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Volume-5, Issue-6 E-ISSN: 2347-2693

Cross-Domain Sentiment Classification Using SST

Megha P.K[1*] and Hima K.G[2]

[1*] Royal College of Engineering and Technology, Thrissur, Kerala
[2] Royal College of Engineering and Technology, Thrissur, Kerala

**Corresponding Author: meghalighi@gmail.com*

Available online at: www.ijcseonline.org

Received: 22/May/2017, Revised: 06/Jun/2017, Accepted: 24/Jun/2017, Published: 30/Jun/2017

Abstract— Sentiment analysis refers to the use of natural language processing and machine learning techniques to identify and extract subjective information in a source material like product reviews. Due to revolutionary development in web technology and social media reviews can span so many different domains that it is difficult to gather annotated training data for all of them. Across domain sentiment analysis invokes adaptation of learned information of some (labeled) source domain to unlabelled target domain. The method proposed in this project uses an automatically created sentiment sensitive thesaurus(SST) for domain adaptation. Based on the survey conducted on related literature, we identified L1 regularized logistic regression is a good binary classifier for our area of interest.This makes our project more accurate in sentiment classification.We can use this as an application for product reviews. In addition to the previous work we propose the use of senti wordnet and adjective adverb combinations for those effective feature learning.

Keywords- *Sentiment Analysis, SST*

I. INTRODUCTION

Cross domain sentiment classification is a method of classifying the sentiments as positive or negative.Sentiment analysis refers to the use of text analysis for extracting subjective information. Sentiment classification has been applied in numerous tasks such as opinion mining, opinion summarization, contextual advertising and market analysis .Sentiments in this system refers to reviews in various domains. Users express opinions about products or services they consume in blog posts, shopping sites, or review sites. It is useful for both consumers as well as for producers to know what general public think about a particular product or service. Automatic document level sentiment classification is the task of classifying a given review with respect to the sentiment expressed by the author of the review. For example, a sentiment classifier might classify a user review about a movie as positive or negative depending on the sentiment expressed in the review. We define cross domain sentiment classification as the problem of learning a binary classifier (i.e. positive or negative sentiment) given a small set of labeled data for the source domain, and unlabeled data for both source and target domains. In particular, no labeled data is provided for the target domain. In this proposed system, we describe a cross-domain sentiment classification method. In this work, the lexical elements (unigram or bigram) in a review are taken and score of each lexical elements is calculated using Senti wordnet. Using this trained dataset is created and the test data will be classified according to this trained data [3]. In this work, logistic regression based algorithm is used for sentiment classification.

In previous work,various methods have been used for classification in single domain.Some of the classification method involves Bayesian classification,Entropy based method,Support Vector machine,Structural correspondance learning etc.They are desribed as follows:In Domain Adaptation with Structural Correspondance Learning: Structural Correspondence Learning is one of the first algorithm for domain adaptation. Many NLP tasks suffers lack of training data in the domain. To face this challenge the possible solution is adapt a source domain (known domain) to a target domain (new domain). This is called Domain Adaptation[4]. Structural correspondence learning (SCL) is a general technique (a Domain adaptation algorithm) which can be applied to feature based classifiers, proposed by Blitzer[4].The key idea of SCL is to identify correspondences among features from different domains by modeling their correlations with pivot features. Pivot features are features which behave in the same way for discriminative learning in both domains. Structural correspondence learning involves a source domain and a target domain. Both domains have ample unlabeled data, but only the source has labeled training data. The SCL algorithm involves selection of pivot features, training a binary classifier for every pivot features. The simplest criterion for selecting pivot feature is that it

should occur frequently in the unlabeled data of both domains. The binary classifier here acts as prediction function. These binary classification problems can be trained from the unlabeled data, since they merely represent properties of the input. If the features are represented as a binary vector x, these can be solved by using m linear predictors. Since each instance contains features which are totally predictive of the pivot feature , we never use these features when making the binary prediction. That is, we do not use any feature derived from the right word when solving right token pivot predictor. Then arrange the pivot predictor weight vectors in matrix W. Apply Singular Value Decomposition to W, and select the h top left singular vectors . Train a new model on the source data augmented with x. Singular Value Decomposition (SVD) decompose a matrix A of order m X n, into product three matrices: A= L S Transpose(V), where L is an orthonormalized matrix of order m X m , S is a diagonal matrix of order m X n and V is the orthogonal matrix of order n X n.

In Sentiment Classification Using Machine Learning Techniques: This work [2] mainly examine the effectiveness of applying machine learning techniques to the sentiment classification problem. A challenging aspect of this problem that seems to distinguish it from traditional topic-based classification is that what topics are often identifiable by keywords alone, sentiment can be expressed in a more subtle manner. Sentiment classification would be helpful in business intelligence applications and recommender systems, where user input and feedback could be quickly summarized. The main aim of this work was to examine whether it suffices to treat sentiment classification simply as a special case of topic based categorization with the two topics being positive sentiment and negative sentiment, or whether special sentiment categorization methods need to be developed. Three basic standard algorithms , Naive Bayes Classification, Maximum Entropy Classification, Support Vector Machine are experimented in this work.

The reminder of the paper is organized as follows: section II presents about Analysis,Section III deals with the Design And Implementation, section IV shows the observations. Section V gives the concluding remarks and finally the paper ends with few references.

II. DESIGN AND IMPLEMENTATION

Cross Domain Sentiment Classification is a method of classification applied when we do not have any labeled data for a target domain but have some labeled data for multiple other domains, designated as the source domain. It focuses on the challenge of training a classifier from one or more domains (source domains) and applying the trained classifier in a different domain (target domain).A cross-domain sentiment classification system must overcome two main challenge.First, it must identify which source domain features are related to which target domain features. Second, it requires a learning framework to incorporate the information regarding the relatedness of source and target domain features.

A. Input Design

We use labeled data from multiple source domains and unlabeled data from source and target domains to represent the distribution of features.Our first Step is,Given a labeled or an unlabeled review, we first split the review into individual sentences.This is done in the Preprocessing Stage of our process. The review given will be the input to the first stage. On moving to the next stage or the attained pos tagged words are fetched to this stage,hence to the input given to sentiWordNet will be the pos tagged sentences .The score obtained from this stage will be the input for the next stage,that is the input given to the logistic regression will be the score calculated using sentiwordnet.

B. Output Design

After preprocessing stage of our execution we got the output in the form of POS tagged sentences. this output is given to the next stage as the input,and the next is the sentiWordNet,at this stage the score is calculated and this score will be the output.this score is given to the next logistic regression and the output of this stage will be the prediction that the given sentences or review is positive or not.

C. Module Description

We describe a sentiment classification method that is applicable when we do not have any labeled data for a target domain but have some labeled data for multiple other domains, designated as the source domains. In this our project is mainly divided into four modules,

1. Preprocessing: In this module,First, we select the lexical elements that co-occur with in a review sentence as features. Second, from each source domain labeled review sentence in which the sentence occurs, we create sentiment features by appending the label of the review to each lexical element we generate from that review.we use the notation *P to indicate positive sentiment features and *N to indicate negative sentiment features.In addition to word-level sentiment features,we replace words with their POS tags to create POS-level sentiment features. POS tags generalize the word-level sentiment features, thereby reducing feature sparseness.We then apply a simple word filter based on POS tags to select content words (nouns, verbs, adjectives, and adverbs).

The preprocessing stage of a sentence is described in figure 3.1

sentence	Excellent and broad survey of the development of civilization.
POS tags	Excellent/JJ and/CC broad/JJ survey/NN1 of/IO the/AT development/NN1 of/IO civilization/ NN1
lexical elements (unigrams)	excellent, broad, survey, development, civilization
lexical elements (bigrams)	excellent+broad, broad+survey, survey+development development+ civilization
sentiment features (lemma)	excellent*P , broad*P, survey*P, excellent+broad*P, broad+survey*P
sentiment features (POS)	JJ*P, NN1*P, JJ+NN1*P

TABLE I

GENERATING LEXICAL ELEMENTS AND SENTIMENT FEATURES.

2. SentiWordNet:SentiWordNet is a lexical resource for opinion mining [1]. SentiWordNet assigns to each synset of WordNet two sentiment scores: positivity, negativity..SentiWordnet is an online dictionary and it provides positive and negative score for each lexical elements.
3. Logistic Regression:In this module,we give the input obtained from the previous module that means the score obtained using the SentiWordNet and the labelled elements are given as the inputs.And in this module this input will become the trained set.And when we give the test data it could predict whether it is positive or not.
4. Cross Domain:Till now, the review from a single domain is classified.Now the classification is extended for multi-ple domains.

D. Implementation

We use labeled data from multiple source domains and unlabeled data from source and target domains to repre-sent the distribution of features.Our first Step is,Given a labeled or an unlabeled review, we first split the review into individual sentences.This is done in the Preprocessing Stage of our process. The review given will be the input to the first stage.,First, we select other lexical elements that co-occur with in a review sentence as features. Second, from each source domain labelled review sentence in which the sentence occurs, we create sentiment features by appending the label of the review to each lexical element we generate from that review.we use the notation *P to indicate positive sentiment features and *N to indicate negative sentiment features. In addition to word-level sentiment features,we replace words with their POS tags to create POS-level sentiment features. POS tags generalize the word-level sentiment features, thereby reducing feature sparseness.We then apply a simple word filter based on POS tags to select content words (nouns, verbs, adjectives, and adverbs).After preprocessing stage of our execution we got the output in the form of POS tagged sentences. this output is given to the next stage as the input.On moving to the next stage or the attained pos tagged words are fetched to this stage,hence to the input given to sentiWordNet will be the pos tagged sentences .The score obtained from this stage will be the input for the next stage. SentiWordNet is a lexical resource for opinion mining. SentiWordNet assigns to each synset of WordNet two sentiment scores: positivity, negativity.SentiWordnet is an online dictionary and it provides positive and negative score for each lexical elements. at this stage the score is calculated and this score will be the output.The score obtained from this stage will be the input for the next stage,that is the input given to the logistic regression will be the score calculated using sentiwordnet. we give the input obtained from the previous module that means the score obtained using the SentiWordNet and the labelled elements are given as the inputs.And in this module this input will become the trained set.And when we give the test data it could predict whether it is positive or not.the implementation is only done in a single domain,now this is to implement in cross domain.

III. RESULT ANALYSIS

The proposed system is divided into four modules.Out of this three modules are completed and this three modules are individually tested for finding errors and to improve performance.These testing are carried at the programming stage itself.In unit testing,a test is performed on whether the data is correctly splitted into sentences and whether the POS tags was correct or not.Also testing is done to check whether the data is labelled correctly. In integrated testing,the three modules are combined ie,after the preprocessing of the text documents it is given to the sentiwordnet for calcu-lating the score of each lexical elements.Dataset is then created.These two modules are combined with the logistic regression algorithm for classifying the unlabelled data. The testing was done on this combined modules to check for errors.After performing the integration test-ing,a careful output testing is done to check whether the correct POS tags was generated and whether it classifies the review correctly.Output is tested for various inputs and checks the system behave properly for each input given.

Two different classification algorithms are taken for doing the experiments.Using Bayesian classification method, the same reviews are classified and we get an accuracy of.Using logistic regression technique,we get an accuracy of. Thus it is concluded that logistic regression method proves to be more accurate than Bayesian classification.

IV. CONCLUSION

Sentiment analysis is found to be the method of classifying huge reviews by analysing its opinion strength. We have implemented a system that classifies the reviews from a single domain.L1 logistic regression method is used for this classification. Furthermore we have to implement this classifier in multiple domains in future and make a GUI interface.

REFERENCES

[1] Andrea Esuli , Fabrizio Sebastiani, " *SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining*", Proc. of the 5th Conf. on Language Resources and Evaluation (LREC06), 2006.

[2] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan."*Thumbs up? sentiment classification using machine learning techniques*". In EMNLP 2002, pages 79-86.

[3] Danushka Bolegalla ,David Weir, John Caroll "*Using Multiple Sources to Construct a Sentiment Sensitive Thesaurus for Cross-Domain Sentiment Classification*".

[4] John Blitzer, Ryan McDonald, and Fernando Pereira."*Domain adaptation with structural correspondence learning*". In EMNLP 2006.

[5] Farah Benamara, Sabatier Irit, Carmine Cesarano, Napoli Federico, Diego Reforgiato, " *Sentiment Analysis: Adjectives and Adverbs are better than Adjectives Alone* ", In Proc of Int Conf on Weblogs and Social Media , 2007.

Authors Profile

Megha P.K. pursued Bachelor of Technology from Calicut University, in 2014. She is currently pursuing Master of Technology under APJ Abdul Kalam Tecnological University, Kerala, India. Her main research work focuses on Natural Language Processing.

Hima K.G received her M.Tech. Degree in Computer Science And engineering and B.Tech. Degree in Computer Science . He is currently working as Assistant Professor at Royal College of Engineering and Technology, Thrissur, Kerala, INDIA.She has 5 years of teaching experience . He has published 4 research papers in various reputed international journals. Her main research interest is in Image processing.

.

