Merged Agreement Algorithms for Domain Independent Sentiment Analysis

Dinko Lambov, Sebastião Pais, Gaël Dias
University of Beira Interior
{dinko, sebastiao, ddg}@hultig.di.ubi.pt

Abstract

In this paper, we consider the problem of building models that have high sentiment classification accuracy across domains. For that purpose, we present and evaluate three new algorithms based on multi-view learning using both high-level and low-level views. Our experimental results present accuracy levels of 80% with two views, combining SVM classifiers over high-level features and unigrams compared to 77.1% for the state-of-the-art SAR algorithm [7].

1 Introduction

Over the past few years, there has been an increasing number of publications focused on the classification of sentiment in texts. However, as stated in [1, 6, 2, 4], most research have focused on the construction of models within particular domains and have shown difficulties to cross thematic spheres. Within this context, three main approaches have been tackled. The first one is to train a classifier on a domain-mixed set of data as in [1, 4]. The second solution is to propose high-level features, which do not depend so much on topics such as part-of-speech statistics or other semantic resources as in [6, 8]. In this case, high level representations do not reflect the topic of the document, but rather the text genre. The third solution is to propose multi-view learning algorithms. The basic idea is to train at least two classifiers on one source domain and then update the set of labeled examples with new examples from a target domain when both classifiers agree on the class of the unlabeled examples. This process is then iterated until convergence. Within this context, [7] proposed a new algorithm for probabilistic multi-view learning, which uses the idea of stochastic agreement between views as regularization, called Stochastic Agreement Regularization (SAR). In parallel, [11] proposed a multi-view learning approach to improve the classification accuracy of polarity identification of Chinese product reviews based on translated English reviews. For that purpose, they use the co-training algorithm [3] with an agreement constraint combined with two SVM classifiers. In this paper, we propose to compare the SAR algorithm [7] with the co-training algorithm strategy constrained by agreement (i.e. a multi-view learning paradigm) as in [11]. Within this context, we propose three multi-view learning algorithms, which best one presents accuracy levels across domains of 80% compared to 77.1% for the SAR.

2 Characterizing Subjectivity

In many works [10], low-level features (e.g. unigrams and bigrams) have been used to characterize subjectivity. In order to cross domains, [8] proposed to use high-level features, which are statistically relevant for sentiment classification. For that purpose, we used their 7 best features: proportions of affective words, semantically oriented adjectives, dynamic adjectives, conjecture verbs, marvel verbs, see verbs and the level of abstraction of nouns. For the unigram model, we used TF.IDF weights for all the lemmas withdrawing stop words. In all our experiments, one view is based on high-level features while the second view is based on low-level features, in particular unigrams as experiments with bigrams did not evidenced improved results.

3 Subjective/Objective Text Data Sets

To perform our experiments, we used three manually annotated standard corpora and built one larger corpus based on Web resources, which texts were automatically annotated as objective or subjective so that manual work was leveraged. The first data set is based on the Multi-Perspective Question Answering (MPQA) Opinion Corpus[1]. Following the idea of [9] who propose to classify texts based only on their subjective/objective parts, we built a corpus of 100 objective (resp. subjective) texts by randomly selecting sentences containing only subjective or objective phrases. The second corpus (RIMDB) is the subjectivity data set v1.0[2], which contains 5000 subjective and

[1] http://www.cs.pitt.edu/mpqa/

[2] http://www.cs.cornell.edu/People/pabo/movie-review-data/

5000 objective sentences collected from movie reviews data [9]. Similarly to the MPQA corpus, we built a corpus of 100 objective (resp. subjective) texts by randomly selecting only subjective or objective sentences. The third corpus (CHES) was developed by [5] who manually annotated a data set of objective and subjective documents[3] from newspapers.

However, the MPQA, RIMDB and CHES are small collections of subjectivity labeled documents. Moreover, their construction is labor intensive as they have to be manually tagged by domain experts. To avoid this problem, we propose to compare Wikipedia and Weblogs texts to reference objective and subjective corpora and show that Wikipedia texts (resp. Weblogs) are representative of objectivity (resp. subjectivity) based on language modeling.

For that purpose, we downloaded part of the static Wikipedia dump archive[4] and automatically crawled Weblogs from different domains to gather two huge collections of texts. We then built two language models from these two data sets. So, our experiment aims at showing that the subjective part of the RIMDB[5] should be more probable than the objective part for the subjective language model (i.e. Weblogs), and vice and versa. This probability is transformed into perplexity (Px) and entropy (H) measures within the CMU-Toolkit[6]. The results are given in Table 1 for a trigram language model.

Table 1. Language modeling.

	Wikipedia	Weblogs
Obj.	Px=691.27 - H=9.43	Px=2027.06 - H=10.99
Subj.	Px=880.67 - H=9.75	Px=1991.09 - H=9.75

To summarize the results, the trained model Wikipedia shows lower perplexity and entropy for the objective sentences than for the subjective sentences. The opposite happens when using the trained model Weblogs. As a consequence, our assumptions are confirmed as objective (resp. subjective) sentences are intrinsically closer to the Wikipedia (resp. Weblogs) model than subjective (resp. objective) ones. Based on this analysis, we are able to automatically build a subjective/objective data set of learning examples based on common sense judgments. For this study, we built the fourth corpus (WBLOG) by randomly selecting 100 objective and 100 subjective texts, respectively from all Wikipedia texts and all Weblogs.

4 Multi-View Learning

While semi-supervised learning is usually associated to small labeled data sets and tries to automatically increase the number of labeled examples, multi-view learning aims at learning a compromise model of different views. A classical semi-supervised algorithm is the well-known co-training algorithm [3]. [11] proposed a slight modification of the co-training algorithm by introducing an agreement constraint, which can be thought as a way of providing a multi-view learning approach. Within this context, new labeled examples are included in the set of labeled examples if all classifiers agree on their labels. As such, all classifiers tend to converge to a compromise learner. This is the definition of the multi-view paradigm. Different works have been proposed following this approach, but SAR [7] is certainly the best reference up-to-date for sentiment classification.

4.1 The SAR Algorithm

[7] proposed the Stochastic Agreement Regularization (SAR) algorithm to deal with polarity cross-domain classification. In particular, SAR models a probabilistic agreement framework based on minimizing the Bhattacharyya distance between models trained using two different views. It regularizes the models from each view by constraining the amount by which it allows them to disagree on unlabeled instances from a theoretical model. Their co-regularized objective, which has to be minimized, is defined in Equation 1 where L_i for $i = 1..2$ are the standard regularized loglikelihood losses of the probabilistic models p_1 and p_2, $E_u[B(p_1(\theta_1), p_2(\theta_1))]$ is the expected Bhattacharyya distance between the predictions of both models on the unlabeled data, and c is the weighting constant of the agreement.

$$\text{Min } L_1(\theta_1) + L_2(\theta_2) + cE_u[B(p_1(\theta_1), p_2(\theta_1))]. \quad (1)$$

4.2 Merged Agreement Algorithms

In this section, we propose three algorithms based on the well-known co-training by introducing different agreement constraints. On the one hand, we know that high-level features provide strong opinion evidence across domains [8]. On the other hand, word-based models show remarkable results for in-domain classification tasks [9]. As a consequence, we expect that agreement between low-level classifiers and high-level classifiers may allow to self-adapt to new domains.

4.2.1 The MAA and BMAA Algorithms

The Merged Agreement Algorithm (MAA) is an adaptation of the algorithm proposed in [11]. It is based on the co-training algorithm with agreement, but instead of just taking into account unlabeled examples with similar predictions from both classifiers to update the set of labeled examples such as in [11], we impose that only the examples with highest confidence upon agreement are added to the labeled list. Basically, the MAA takes two main inputs: a set of labeled

[3] http://www.tc.umn.edu/∼ches0045/data/

[4] http://download.wikimedia.org/enwiki/

[5] Experiments were run over all corpora and evidenced similar results.

[6] http://www.speech.cs.cmu.edu/SLM_info.html

examples from one domain (L), the source domain, and a set of unlabeled examples from another domain (U), the target domain. After training on L, both classifiers classify unlabeled documents from U. If both classifiers agree on their predictions, the unlabeled documents are added to an agree list for each classifier with the categorization label and the classification confidence. Finally, the P positive (subjective) and N negative (objective) documents with higher confidence values are transferred from the U to L.

It is important to point at the fact that the MAA algorithm may produce unbalanced data sets. Indeed, from both agree lists, we may update the labeled list with more positive examples than negative ones and vice and versa as classifiers may agree more on one class than another. As a consequence, we propose to modify the MAA algorithm so that it balances the parameter values P and N at each iteration. So, if the number of predicted subjective or objective documents is equal to 0, it is used as a stopping criterion. Otherwise, the minimum number of positive or negative new labeled examples is chosen to update L. We call this method the Balanced Merged Agreement Algorithm (BMAA), our second algorithm proposal.

4.2.2 The BMAADR Algorithm

With the MAA and the BMAA, the most confidently predicted P and N examples are selected to update L. However, we may update L with a positive example from one classifier $H1$, which agrees on the classification of the second classifier $H2$ but where the difference between each confidence is high. As a consequence, we may update L with examples where only one of the classifiers is very confident about the classification although they agree on the classification. The idea of this proposal is to measure an "average" confidence value for all examples for which there is agreement between classifiers so that the highest "on average" new labeled examples are added to L. For that purpose, after each classification on U, both agree lists are sorted by decreasing classification confidence. So, each document is located at one position in each agree list and we reckon a new position, which is the average of the positions of the document d in both lists. Finally, we sort the documents according to their new position and the best P positive and N negative examples are added to L keeping balance between data sets. This method is described in Algorithm 1 and called the Balanced Merged Agreement Algorithm Using Documents Rank (BMAADR).

Algorithm 1 The BMAADR Algorithm.

Input: L a set of labeled examples from one domain, U a set of unlabeled examples from another domain, $P = N = X$
Output: Trained classifier $H2$
$H1.AgreeList \leftarrow \{\}$
$H2.AgreeList \leftarrow \{\}$
for k iterations **do**
  Train a classifier $H1$ on view $V1$ of L
  Train a classifier $H2$ on view $V2$ of L
  Allow $H1$ and $H2$ to label U
  for all $d \in U$ **do**
    if $H1.Class[d] = H2.Class[d]$ **then**
      $H1.AgreeList \leftarrow AgreeList \cup \{< d; H1.Class[d] >\}$
      $H2.AgreeList \leftarrow AgreeList \cup \{< d; H2.Class[d] >\}$
    end if
  end for
  $Sort(H1.AgreeList, byDecrConf.)$
  $Sort(H2.AgreeList, byDecrConf.)$
  for all $d \in H1.AgreeList$ **do**
    $Rank_d = \frac{\sum_{i \in \{H1,H2\}} Rank_d^{i.AgreeList}}{2}$
    $topAgreeList \leftarrow (d, Rank_d)$
  end for
  $L \leftarrow L \cup \{$Balanced P positive and N negative examples with the lower rank from $topAgreeList\}$
end for

5 Experiments and Results

In this section, we present the results obtained by using the multi-view learning techniques combining high-level and low-level features. First, we will present the results obtained by the SAR algorithm, which will form the baseline and then compare these results with those obtained with the proposed MAA, BMAA and BMAADR algorithms. All experiments are performed on a leave-one-out 5 cross validation basis with SVM classifiers. In particular, we use the SVMlight package[7] for classification and the MontyTagger of the MontyLingua package[8] for part-of-speech tagging. To perform the experiments with SAR, we used two views generated from a random split of unigrams together with maximum entropy classifiers with a unit variance Gaussian prior. Indeed, the actual implementation of SAR does not allow to test it with different views but only with random subsets of views. To perform the experiments for MAA, BMAA and BMAADR algorithms, the first view is based on the seven high-level features expressed in section 2 and the second view, with the set of unigrams. As a consequence, we expect that the low-level classifier will gain from the agreements with the high-level classifier and will self-adapt to different new domains. The results are presented in Table 2. In order to test models across domains, we propose to train different models based on only one do-

[7] http://svmlight.joachims.org/
[8] http://web.media.mit.edu/ hugo/montylingua/

main at each time and test the resulting unigram classifier over all other domains together. So, each percentage in Table 2 can be expressed as the average results over all data sets.

Table 2. Accuracy results for unigrams in %.

	MPQA	RIMDB	CHES	WBLOG
SAR	63.7	77.1	72.3	59.7
MAA	59.1	63.5	75.6	69.4
BMAA	59.4	65.2	79.5	69.7
BMAADR	59.4	65.4	80.0	69.9

As results differ at each iteration of the algorithm, we found important to illustrate the behavior of each classifier in Figure 1 in terms of accuracy along the different iterations. The MAA classifier improves its accuracy just in the first few iterations and then starts to loose in accuracy. This is mainly due to the fact that the unbalanced labeled examples impair the performance. In this case, the average accuracy across domains reaches 75.6% in the best case, which is worse than the SAR best performance of 77.1%. However, the BMAA and BMAADR show a different behavior as their accuracy decreases slightly between the sixth and eighth iterations and then remains almost constant for both classifiers. As such, they benefit of the agreement of both classifiers in the first iterations. The best accuracy is obtained by the BMAADR algorithm, which reaches an average accuracy of 80%, which outperforms SAR. Moreover, we see that automatically building a labeled data set, such as the WBLOG, can lead to interesting results as it shows the second best performance. The obtained results also show that SAR performs better in the cases of exclusively objective and subjective data sets (RIMDB and MPQA), while in the case of the other two data sets annotated at document level, the best classification accuracies are obtained by the BMAADR. As a consequence, we can say that the BMAADR algorithm is the best performing algorithm for real-world texts situations.

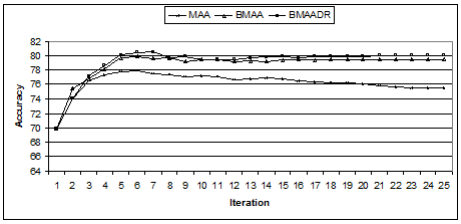

Figure 1. Accuracies of Merged Algorithms.

6 Conclusions

In this paper, we proposed to use a multi-view approach to address the problem of cross-domain sentiment classification. For that purpose, we presented three different algorithms based on an adaptation of the co-training by introducing different agreement constraints following the idea of [11]. The results showed the effectiveness of the proposed approach by combining high-level and low-level features as two different views. In particular, the best results showed accuracy of 80% across domains with BMAADR compared to 77.1% for SAR, the reference in the field.

References

[1] A. Aue and M. Gamon. Customizing sentiment classifiers to new domains: a case study. In *Proceedings of International Conference on Recent Advances in Natural Language Processing (RANLP 2005)*, pages 207–218, 2005.

[2] J. Blitzer, M. Dredze, and F. Pereira. Biographies, bollywood, boomboxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007)*, pages 187–205, 2007.

[3] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory (COLT 1998)*, pages 92–100, 1998.

[4] E. Boiy, P. Hens, K. Deschacht, and M.-F. Moens. Automatic sentiment analysis of on-line text. In *Proceedings of the 11th International Conference on Electronic Publishing (ELPUB 2007)*, pages 349–360, 2007.

[5] P. Chesley, B. Vincent, L. Xu, and R. Srihari. Using verbs and adjectives to automatically classify blog sentiment. In *Proceedings of the AAAI Symposium on Computational Approaches to Analyzing Weblogs (AAAI/CAAW 2006)*, pages 27–29, 2006.

[6] A. Finn and N. Kushmerick. Learning to classify documents according to genre. *American Society for Information Science and Technology, Special issue on Computational Analysis of Style*, 57(11):1506–1518, 2006.

[7] K. Ganchev, J. Graca, J. Blitzer, and B. Taskar. Multi-view learning over structured and non-identical outputs. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence (UAI 2008)*, pages 204–211, 2008.

[8] D. Lambov, G. Dias, and V. Noncheva. Sentiment classification across domains. In *14th Portuguese Conference on Artificial Intelligence (EPIA 2009)*, 2009.

[9] B. Pang and L. Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004)*, pages 271–278, 2004.

[10] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 79–86, 2002.

[11] X. Wan. Co-training for cross-lingual sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing (ACL/IJCNLP 2009)*, pages 235–243, 2009.