Transverse Subjectivity Classification

Dinko Lambov
University of Beira Interior
HULTIG/CMAT
Covilhã, Portugal
dinko@hultig.di.ubi.pt

Gaël Dias
University of Caen Basse-Normandie
GREYC/UCBN/ENSICAEN
Caen, France
gael.dias@unicaen.fr

ABSTRACT

In this paper, we consider the problem of building models that have high subjectivity classification accuracy across domains. For that purpose, we present and evaluate new methods based on multi-view learning using both high-level (i.e. linguistic features for subjectivity detection) and low-level features (i.e. unigrams and bigrams). In particular, we show that multi-view learning, combining high-level and low-level features with adapted classifiers, can lead to improved results compared to one of the state-of-the-art algorithms called Stochastic Agreement Regularization. In particular, the experiments show that dividing the set of characteristics into three views returns the best results overall with accuracy across domains of 91.3% for the Class-Guided Multi-View Learning Algorithm, which combines both Linear Discriminant Analysis and Support Vector Machines.

Categories and Subject Descriptors

I.2.6 [**Learning**]: Language Acquisition; I.2.7 [**Natural Language Processing**]: Text Analysis

General Terms

Algorithms, Experimentation.

Keywords

Cross-Domain, Subjectivity Learning, Multi-View Learning, Sentiment Analysis.

1. INTRODUCTION

Over the past few years, an increasing number of publications have been focusing on classification of sentiment in texts. However, as first stated in [1, 9, 4], most research has focused on the construction of models within particular domains and have shown difficulties to cross thematic spheres. As a consequence, a great deal of studies have been focusing on cross domain sentiment classification [2, 10, 24, 27, 21, 30, 5, 11, 14].

However, most papers deal with polarity as the essence of subjectivity. But subjectivity can be expressed in different ways as proposed in [20] such as evaluation (positive or negative), specificity (clear or vague), proximity (near or far social proximity), intensity (more or less) and certainty (confident or doubtful). This model is usually called the appraisal model. Moreover, the approaches proposed so far have been tested over the well-known Amazon data set[1] made of reviews of Amazon products gathered by [2] and labelled as either positive or negative. Although this data set is useful for evaluation purposes, a question must be raised about the kind of information learnt: “Do the models learn text polarity for specific text genres or not?”. Indeed, the fact that all texts are product reviews (usually literacy criticism by lay writers) may influence the way polarity is learnt. At the end, cross domain polarity classifiers can be obtained with high accuracy results, but they certainly depend on a particular text genre (i.e. product reviews). As a consequence, in this paper, we will focus on subjectivity classification and not just polarity, and we will test our models on a new data set gathering texts from different domains and genres.

Within cross domain sentiment classification, four main approaches can be distinguished. One possible approach is to train a classifier on a domain-mixed set of data instead of training it on one specific domain as proposed in [1, 4]. Another possibility 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 [9]. A third approach is proposed by [2], also called the sentiment transfer approach, which goal is to find words from different domains indicating the same sentiment. This methodology has recently shown a great impact in the field with works proposed by [24, 21, 30, 5, 14, 11]. In parallel, a fourth framework has also emerged using multi-view learning as in [10, 27], which overall idea is to propose an agreement learning model, which should satisfy both the restrictions of the source domain and the target domain. Theoretically, the last approach shows great potential and will be the main focus of our work.

Thus, in this paper, we introduce different multi-view learning algorithms using both high-level features (e.g. level of affective words, level of abstraction of nouns) and low-level characteristics (e.g. unigrams, bigrams) to learn subjective language across domains. In particular, we will (1) build a new data set for subjectivity learning, (2) propose a new feature for cross domain subjectivity classification, the level of abstraction of nouns and (3) define different multi-view learning algorithms based on the ideas of [10, 27]. We

WISDOM '12 August 12 2012, Beijing, China
Copyright 2012 ACM 978-1-4503-1543-2/12/08 ...$15.00.

[1] http://www.cs.jhu.edu/∼mdreze/datasets/sentiment

will finally show that dividing the set of characteristics into three views and applying the new Class-Guided Multi-View Learning Algorithm (C-GMVLA) lead to maximum performance and outperform the stochastic agreement regularization (SAR) algorithm [10], although simple Co-training [3] is difficult to beat. Within this context, the best overall accuracy result reaches 91.3% by combining Linear Discriminant Analysis (LDA) and Support Vector Machines (SVM) classifiers over our multi-domain multi-genre data set.

2. SENTIMENT MULTI-VIEW LEARNING

Over the past few years, multi-view learning proposals have emerged for cross domain polarity classification. But this approach has recently been neglected in favour of the sentiment transfer approach, which has received great focus by the research community [24, 21, 30, 5, 14, 11]. Nevertheless, we deeply believe that the multi-view learning approach can lead to new insights in the field due to its adapted and well-founded background for the specific task of cross domain subjectivity classification. Within this approach, [10] propose a co-regularization framework for learning across multiple related tasks with different output spaces. They present a new algorithm for probabilistic multi-view learning, which uses the idea of stochastic agreement between views as regularization. Their algorithm called Stochastic Agreement Regularization (SAR) works on structured and unstructured problems and generalizes to partial agreement scenarios. For the full agreement case, their algorithm minimizes the Bhattacharyya distance between the models of each of two views. In parallel, [27] proposes a co-training approach to improve the classification accuracy of polarity identification of Chinese product reviews. First, machine translation services are used to translate English training reviews into Chinese reviews as well as translate Chinese test reviews and additional unlabelled reviews into English reviews. Then, the classification problem can be viewed as two independent views: the Chinese view with only Chinese features and the English view with only English features. Then, they use the Co-training algorithm [3] with an agreement constraint and SVM to make full use of the two redundant views of features. Experimental results show that the proposed approach can outperform inductive and transductive classifiers. Based on these promising results of [10, 27], we propose an exhaustive study of Co-training-based algorithms and compare them with success to SAR, the state-of-the-art multi-view learning algorithm in the field.

3. LEARNING DATA SETS

To perform our experiments, we built four different corpora based on (1) three manually annotated well-known standard corpora and (2) one automatically crawled from web resources (i.e. Wikipedia texts and Weblogs), which can be automatically annotated as objective or subjective as proved in our previous works. As such, we aim to propose a new standard data set for cross domain subjectivity (vs. polarity) classification. The first corpus (MPQA) is based on the Multi-Perspective Question Answering Opinion Corpus[2] [28, 29]. Based on the work done by [22] who classify texts based only on their subjective/objective parts, we built a corpus of 100 objective (resp. subjective) texts by randomly selecting sentences from MPQA containing exclusively subjective or objective phrases. This case represents the "ideal" case where all the sentences in subjective or objective texts are respectively either subjective or objective. The second corpus (RIMDB) is based on the subjectivity dataset v1.0[3], which contains 5000 subjective and 5000 objective sentences collected from movie reviews data [22]. 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 [8] who manually annotated a data set of objective and subjective documents[4]. For the fourth corpus (WBLOG), the latent idea was to compare Wikipedia texts and Weblogs to reference objective and subjective corpora and show that Wikipedia texts are representative of objectivity whereas Weblogs are representative of subjectivity. For that purpose, we proposed in previous works an exhaustive evaluation based on (1) the Rocchio classification method [25] for different part-of-speech tag levels and (2) language modelling. Both results confirmed the initial assumptions that texts from Wikipedia (resp. Weblogs) convey objective (resp. subjective) contents. In order to build our automatically labelled corpus from web resources, we downloaded part of the static Wikipedia dump archive[5] and automatically spidered Weblogs from different domains. As such, we proposed a balanced multi-domain multi-genre data set with 100 objective texts and 100 subjective texts.

4. CHARACTERIZING SUBJECTIVITY

In many works [23], low-level features (i.e. unigrams and bigrams) have been used to characterize subjectivity. However, it is well-known that low-level features show critical capacities to cross domain when directly included as features in the learning process. As a consequence, studies have proposed to use high-level features, which are statistically relevant for subjectivity classification. For that purpose, we used six well-known features and proposed a new feature based on linguistic evidence of subjectivity, the level of abstraction of nouns, which may represent specificity in the appraisal model. The six classical features are: (1) the proportion of affective words in texts using WordNet Affect Lexicon [26], (2) the proportion of semantically oriented adjectives in texts using the set of semantic orientation labels assigned as in [12], (3) the proportions of dynamic adjectives in texts using the set of dynamic adjectives manually identified by [13] and (4-6) the proportion of conjecture, marvel and see verbs in texts using the classification of verbs available in [17].

As most of theses resources have been built on the idea of characterizing subjectivity only based on polarity, we propose a new feature, which evaluates the average degree of generality/specificity of a text using WordNet [19]. Indeed, there is linguistic evidence that level of generality is a characteristic of opinionated texts, i.e. subjectivity is usually expressed in more abstract terms than objectivity. As a consequence, we evaluate the level of abstraction of nouns in a given text by counting the number of paths to the root "entity" for all the nouns in the text, which are contained

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

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

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

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

in WordNet. The higher the average the more objective the text will be and vice versa. As such, we evidence the specificity feature of subjectivity of the appraisal model[6].

Although the use of unigrams and bigrams is insufficient to cross domains, the idea behind multi-view learning is that meaningful (i.e. potential transfer unigrams or bigrams) can be determined by multi-view classifiers trained over a source domain and adjusted to a target domain. For that purpose, we use tf.idf weights for all lemmas withdrawing stop words. In all our experiments, one view will always be based on low-level features, as this model will then be used for tests.

5. MULTI-VIEW ALGORITHMS

Multi-view learning for cross domain subjectivity classification can be addressed following three different ideas. The first approach is based on the idea that a cross domain classifier can be obtained by training two classifiers based on two different views (or feature sets) on a source domain (labelled data set) and tuning it over a target domain (unlabelled data set) by imposing agreement. This is the idea of [10, 27]. The second approach is based on the classical Co-training algorithm proposed by [3], where no agreement is imposed and unlabelled data from the target domain, which are classified with high confidence by any classifier are automatically added to the new training set with the corresponding label for the next iteration of the algorithm. The third approach is based on the idea that at each iteration of the algorithm, the best view (i.e. the best classifier) is used to label the unlabelled data from the target domain and the most confidently classified examples from the selected classifier are added to the new training data set for the next iteration of the algorithm. We call this technique, Guided Multi-View Learning.

5.1 Learning with Agreement

[10] propose the Stochastic Agreement Regularization (SAR) algorithm to deal with polarity cross domain classification. SAR models a probabilistic agreement framework based on minimizing the Bhattacharyya distance [16] between models trained using two different views. It regularizes the models from each view by constraining the amount by which it permits them to disagree on unlabelled instances from a theoretical model.

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

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, $E_u[B(p_1, p_2)]$ is the expected Bhattacharyya distance between the predictions of the two models on the unlabelled data, and c is a constant defining the relative weight of the agreement.

In parallel, [27] proposes a simple adaptation of the Co-training algorithm by imposing an agreement constraint as shown in Algorithm 1. The algorithm is called the Agreement Co-training Algorithm (ACA). It is important to notice that while [10] do not update the initial labelled data set from the source domain, this is not the case for [27], who increases the source data set with confidently classified unlabelled texts from the target domain in a common co-training strategy.

Algorithm 1 The Agreement Co-training.

```
Input: $L$ a set of labelled examples from one domain, $U$ a set of unlabelled examples from another domain
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
            $H1.AgreeList \leftarrow H1.AgreeList \cup \{< d; H1.Class[d] >\}$
            $H2.AgreeList \leftarrow H2.AgreeList \cup \{< d; H2.Class[d] >\}$
        end if
    end for
    $L \leftarrow L \cup$ {the most confidently predicted $P$ positive and $N$ negative examples from $H1$ on $U \in H1.AgreeList$}
    $L \leftarrow L \cup$ {the most confidently predicted $P$ positive and $N$ negative examples from $H2$ on $U \in H2.AgreeList$}
end for
```

One of the main drawbacks of the algorithm proposed by [27] is that it may produce unbalanced data sets and as a consequence bias the learning process. Indeed, from both agree lists of $H1$ and $H2$, we may update the labelled list with more positive examples than negative ones and vice versa, as classifiers may agree more on one class than on another. As a consequence, we propose to modify his algorithm to balance the parameter values of 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 labelled examples is chosen to update the source labelled example list L. This cycle is repeated for k iterations or until there are no positive or negative candidate documents in the agree lists. This method is called the Balanced Agreement Co-training Algorithm (BACA). Another limitation of the algorithm is that different classifiers may agree on the classification of the same unlabelled document but with huge differences in confidence. To avoid this problem, we propose to measure an "average" confidence value for all examples for which there is agreement between classifiers so that the highest "on average" new labelled examples are added to L. For that purpose, after each classification on unlabelled data, both agree lists are sorted by decreasing classification confidence i.e. the best examples are at the top of the agree lists. So, each document is located at one position in the agree list of $H1$ and on another position in the agree list of $H2$. Based on these two ranks in the different sorted agree lists, we reckon a new rank, which is the average of the ranks of the document d in both lists. Finally, we sort the documents according to their new average rank, which is their new confidence value. Then, the best P positive and

[6] It is important to point that calculating the level of abstraction of nouns should be preceded by word sense disambiguation. However, in practice, taking the most common sense of each word gives similar results as taking all the senses on average.

N negative examples are added to the labelled data set L depending on their new confidence value. This method is called the Balanced Agreement Co-training Algorithm Using Documents Rank (BACAUDR) and is directly adapted from Algorithm 1.

5.2 Guided Learning

While all aforementioned algorithms propose an agreement constraint to the learning process, the guided multiview learning paradigm can be seen as a competitive learning process, where at each iteration of a Co-training-like algorithm, the best classifier (i.e. the best view) is chosen to label unlabelled data and update the source domain data set. Within this scope, we propose a new algorithm called the Guided Multi-View Learning Algorithm (GMVLA), which takes three main inputs: a set of labelled examples from one domain (L), the source domain, the set of unlabelled examples from another domain (U), the target domain, and a validation (VL) data set (i.e. a small set of labelled examples from the target domain). The proposed technique uses the validation data set to guide the selection of new training candidates. At the end of each learning iteration, all classifiers are applied to VL and receive an accuracy score. As a consequence, P positive (i.e. subjective texts) and N negative examples (i.e. objective texts) from U with higher confidence values classified by the classifier with best accuracy are added to L. This method is described in Algorithm 2 for two views.

Algorithm 2 The Guided Multi-view Learning.

```
Input: $L$ a set of labelled examples, $U$ a set of unlabelled
  examples, $VL$ a small set of labelled examples from the
  target domain
Output: Trained low-level classifier $H2$
for $k$ iterations do
    Train a classifier $H1$ on view $V1$ of $L$
    Train a classifier $H2$ on view $V2$ of $L$
    Apply $H1$ and $H2$ to $VL$
    if $H1.Acc[VL] > H2.Acc[VL]$ then
        $L \leftarrow L \cup \{$the most confidently predicted $P$ positive
          and $N$ negative examples from $H1$ on $U\}$
    else
        $L \leftarrow L \cup \{$the most confidently predicted $P$ positive
          and $N$ negative examples from $H2$ on $U\}$
    end if
end for
```

Instead of relying only on the global accuracy over the VL data set and choosing the corresponding classifier to guide the learning process, one may choose the classifier with higher precision for subjectivity to label the P positive examples from U and the classifier with higher precision for objectivity to label the N negative examples from U. As such, we can improve the classification problem. Therefore, at each learning iteration, we compare the subjective precision and the objective precision obtained by each classifier over the validation data set VL and choose the best one for each class. Here, unlike the previous method, we expect to reduce the number of wrong examples added to the labelled data set L mainly due to the fact that the best classification accuracy of one classifier may exclusively be due to high accuracy over only subjective or only objective predictions. Indeed, with the GMVLA algorithm, we may label new examples from U as subjective or objective based on the classifier with best accuracy overall although its precision over one of the classes may be low. In this case, our method would take subjective (resp. objective) examples from the best subjective (resp. objective) classifier. This new algorithm is called the Class-Guided Multi-View Learning Algorithm (C-GMVLA) and can be straightforwardly defined from the GMVLA.

6. EXPERIMENTS

In order to evaluate the differences between high-level and low-level features, and in particular to assess the benefits of the new proposed feature (i.e. the level of abstraction of nouns), we first performed a comparative study across domains on our four data sets. For the high-level features, we took into account 7 features (affective words, dynamic and semantically oriented adjectives, conjuncture verbs, see verbs, marvel verbs and level of abstraction of nouns). For the unigram and bigram feature representations, we used all the lemmas[7] inside the corpora withdrawing stop words and weighting lemmas with the classical tf.idf measure. For the cross domain classification task, we proposed to train a model based on one domain only and test the classifier over all the other domains under a leave-one-out 5 cross validation basis for both Linear Discriminant Analysis (LDA)[8] and Support Vector Machines (SVM)[9] classifiers. This procedure is repeated for the four domain corpora MPQA, RIMDB, CHES and WBLOG. As a consequence, the results presented in Table 1 can be expressed as the average results of the classifier trained over a specific domain and tested over all remaining data sets plus 20% of unseen examples of the source domain.

V1 (Algo.)	MPQA	RIMDB	CHES	WBLOG
Uni. (SVM)	58.8%	64.4%	69.9%	63.9%
Bi. (SVM)	57.5%	66.9%	66.5%	62.3%
7F (SVM)	63.1%	70.5%	70.9%	70.2%
7F (LDA)	69.4%	73.5%	73.9%	74.6%
6F (LDA)	67.4%	67.9%	71.6%	73.0%

Table 1: Accuracy results across domain.

Best results overall are obtained for high-level features with the WBLOG corpus as training data set and the LDA classifier with an average accuracy of 74.6%, which means that combining LDA and WBLOG over the seven high-level features to build a cross domain classifier and testing it over the RIMDB, MPQA and CHES corpora on a leave-one-out 5 cross validation basis[10], evidences an average accuracy of 74.6%. The results also show that accuracy drops drastically by learning based on unigrams or bigrams reaching a maximum average accuracy of 69.9% with SVM[11]. In order to evaluate the importance of the level of abstraction of nouns as a clue for subjectivity identification, we proposed to test classification accuracy of the models based on the six

[7] For this task, we used the MontyTagger of the MontyLingua package [18].

[8] The R implementation of LDA was used.

[9] The SVMlight package was used [15].

[10] For each experience, 20% of unseen examples of the source domain are added to the target domain for testing.

[11] The LDA classifier was unable to deal on due time with huge feature sets sizes.

state-of-the-art features[12] without the level of abstraction of nouns and then compared with the full set of seven features. The experimental results clearly show that using the level of abstraction of nouns as a feature leads to improved performance on subjectivity classification tasks for each of the models.

6.1 Results With Agreement

In order to test the agreement multi-view learning approach, we first proposed to test the SAR algorithm [10][13]. In particular, we used two views generated from a random split of low-level features 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 types of views, nor with different classifiers or more than two views[14]. The evaluation process was processed as follows on a leave-one-out 5 cross validation basis. First, we defined a source domain (a labelled data set from one domain corpus) and a target domain (an unlabelled data set from another domain corpus). After training, the low-level classifier was tested over the unseen examples of the source domain plus the unseen examples from the unseen corpora. This operation was repeated four times, each time for a new target domain. For example, we would train the model on the (MPQA,RIMDB) pair, where MPQA is the source domain and RIMDB is the target domain. The model would then be tested on the unseen examples from MPQA, RIMDB, CHES and WBLOG. In fact, this process would be repeated for the following pairs (MPQA,MPQA), (MPQA,CHES) and (MPQA,WBLOG). As such, the results presented in Table 2 are the average accuracies for all four experiments.

V1 and V2	MPQA	RIMDB	CHES	WBLOG
Unigrams	63.7%	77.1%	72.3%	59.7%
Bigrams	59.8%	65.2%	64.9%	62.2%

Table 2: SAR accuracy results.

The results show interesting properties. In particular, models built upon unigrams usually outperform models based on bigrams thus extending to cross domain a situation already evidenced by [23] for in-domain classification. But the most important result is the fact that SAR can improve results compared to single-view classification using high-level features from 74.6% to 77.1%, by just looking at unigrams. This result is particularly interesting as it may show that combining high-level features with low-level features on an agreement multi-view learning paradigm may improve performance. So we performed the same experiments for the ACA [27] and its adapted algorithms, BACA and BACAUDR. In this case, the first view will contain the seven high-level features and the second view will be the set of unigrams or bigrams. Morevoer, the classifier used for the first view is the LDA and SVM for the second view. 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. In Table 3, we show the results obtained using unigrams as low-level features and in Table 4, the results using bigrams for ACA, BACA and BACAUDR.

Algorithm	MPQA	RIMDB	CHES	WBLOG
ACA	59.1%	63.5%	75.6%	69.4%
BACA	59.4%	65.2%	79.5%	69.7%
BACAUDR	59.4%	65.4%	80.0%	69.9%

Table 3: Accuracy Results for Unigrams.

The results show that SAR performs better in the cases of exclusively objective and subjective data sets (i.e. RIMDB and MPQA), while in the case of the other two data sets annotated at document level (i.e. texts do not contain exclusively objective or subjective sentences), the best classification accuracy is obtained by the BACAUDR with 80.0% combining unigrams and seven high-level features with CHES as the source domain. As a consequence, we can say that the BACAUDR algorithm is the best performing algorithm for real-world texts situations. However, some comments must be made. In the proposed method, we rely on the assumption that the domain-independent view based on high-level features restricts the addition of wrongly predicted labels by both classifiers.

Algorithm	MPQA	RIMDB	CHES	WBLOG
ACA	57.5%	69.9%	71.6%	64.7%
BACA	57.5%	76.6%	77.9%	65.2%
BACAUDR	57.5%	76.7%	77.2%	65.5%

Table 4: Accuracy Results for Bigrams.

However, this method suffers from the weakness of the low-level classifier in its initial states, as wrong classifications may lead to produce small sets of examples, which may join the agree lists. Moreover, when both classifiers agree, they do not learn much more, especially if they agree with high-level of confidence in both classifiers. As a consequence, the accuracy is almost constant for the model just after a few iterations. As a consequence, we propose to relax the agreement constraint and evaluate the original Co-training algorithm as well as the proposed two guided algorithms, GMVLA and C-GMVLA.

6.2 Results With Guided

In the first part of this section, we present the results obtained by using the GMVLA and C-GMVLA algorithms based on two views and then compare them to the results obtained by the standard Co-training approach [3]. The experimental set-ups are exactly the same as the ones presented in the previous section. In Table 5, we show the results obtained using set of unigrams as a second view and in Table 6, the ones by using set of bigrams as a second view. Surprisingly, the best overall result is obtained by the combination of high-level features and unigrams trained over the CHES source domain for the Co-training algorithm with 91% accuracy. On the one hand, better results were expected compared to the agreement approach by relaxing this constraint. As such, the overall methodology increases 11% in accuracy for the best results. However, on the other hand, we would have expected better results from the GMVLA and the C-GMVLA algorithms.

[12] The 6 features line (6F) means that the level of abstraction of nouns was omitted from the seven original high-level features.

[13] We must thank Kuzman Ganchev and João Graça for affording us the code of SAR.

[14] This issue will be discussed in the final conclusions.

Algorithm	MPQA	RIMDB	CHES	WBLOG
GMVLA	82.7%	83.0%	90.3%	85.0%
C-GMVLA	82.9%	82.8%	90.4%	85.6%
CO-TRAIN	82.5%	80.6%	91.0%	85.4%

Table 5: Accuracy results with unigrams.

Algorithm	MPQA	RIMDB	CHES	WBLOG
GMVLA	57.5%	76.7%	81.8%	77.6%
C-GMVLA	57.5%	75.7%	81.7%	76.0%
CO-TRAIN	57.5%	76.8%	81.1%	76.1%

Table 6: Accuracy results with bigrams.

In order to better understand the behaviour of both GMVLA and C-GMVLA algorithms[15], we present the accuracy results over the validation data set at each iteration of the learning process in Figure 1.

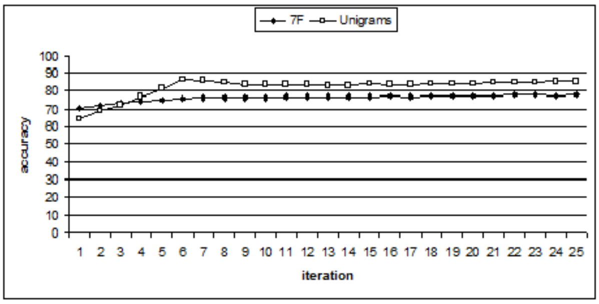

Figure 1: 7 features and unigrams accuracies.

While the accuracy of the classifier based on high-level features (LDA) remains steady, iteration after iteration, the accuracy of the classifier based on low-level features (SVM) steadily improves its accuracy based on the correct guesses of the high-level classifier at the beginning of the learning process[16]. It is clear here that the high-level view "guides" the low-level view classifier so that best performance is obtained. However, this guidance is limited to the first three iterations, as the algorithm will always use the unigram model from iteration 4 until convergence to classify new unlabelled examples as it obtains better accuracy results for this interval. This result can explain the unexpected results obtained by the GMVLA and the C-GMVLA algorithms compared to Co-training, as the guidance is limited to a very few steps and as such classification is almost based only on one classifier, while with the Co-training, the model always integrates eventually well-labelled examples from both classifiers. To solve this problem and focusing on the statistical analysis of high-level features, we discovered that different sets of features have different classification capabilities. As a consequence, we proposed a 3-views methodology to take the most of each feature specificity[17].

7. RESULTS WITH 3-VIEWS GUIDED

In this section, we present the results of the Co-training, the GMVLA and the C-GMVLA for 3 views. In fact, we propose to combine feature classes in a way in which each view will be based on a different type of linguistic information or word representation: (1) bag-of-words representation (i.e. unigrams or bigrams), (2) semantic information (i.e. adjectives, affective words and verbs) and (3) conceptual expression of subjectivity (i.e. level of abstraction of nouns). The idea of using three different views is that different features may have different weights in different domains. For example, we showed in previous works that affective words are strong predictors of subjectivity/objectivity in the news domain, adjectives in Weblogs and verbs in movie reviews. Moreover, the level of abstraction of nouns is a good predictor over domains. As a consequence, we hope that dividing the high-level feature set into different subsets (or views) can show result improvements. In Tables 7, 8 and 9, we show the results obtained using the Co-training, the GMVLA and the C-GMVLA algorithms respectively, using the previous experimental set-ups[18].

V1	V2	V3	MPQA	RIMDB	CHES	WBLOG
Adj.	Uni.	LA	81.5%	84.0%	88.6%	85.4%
Aff.	Uni.	LA	75.6%	87.8%	88.6%	85.2%
Verb	Uni.	LA	78.1%	84.0%	87.1%	85.1%

Table 7: Co-training accuracies with 3 views.

V1	V2	V3	MPQA	RIMDB	CHES	WBLOG
Adj.	Uni.	LA	83.6%	83.0%	89.4%	84.7%
Aff.	Uni.	LA	78.6%	80.7%	90.7%	84.6%
Verb	Uni.	LA	81.2%	82.3%	87.5%	83.5%

Table 8: GMVLA accuracies with 3 views.

V1	V2	V3	MPQA	RIMDB	CHES	WBLOG
Adj.	Uni.	LA	76.6%	83.1%	90.2%	85.2%
Aff.	Uni.	LA	77.8%	83.0%	91.3%	84.8%
Verb	Uni.	LA	75.8%	83.5%	88.6%	84.9%

Table 9: C-GMVLA accuracies with 3 views.

As expected, better results were obtained by using 3 views except for the Co-training algorithm, which tends to introduce noisy examples to the source data set due to the multiplication of views. On the contrary, C-GMVLA takes the best of each feature sets alone. As a consequence, the best result overall is obtained by the combination of affective words, unigrams and the level of abstraction of nouns trained over the CHES source domain for the C-GMVLA algorithm. In this case, the average accuracy across domains reaches 91.3% overtaking the best performance of 91% (resp. 88.6%) of the Co-training algorithm for 2 (resp. 3) views.

[15] The behaviour of both algorithms is the same. As a consequence, we only present the one of the GMVLA algorithm.

[16] In all our experiments, we used $P = N = 2$.

[17] To our knowledge, we are the first to propose a sentiment classification methodology for more than two views.

[18] As bigrams always showed worst results, we discarded them from our result Tables.

Although, the results of the C-GMVLA for 3 views and Co-training for 2 views are near, they can be statistically differentiated based on the Wilcoxon rank-sum test. Results of the p-value are shown in Table 10 and evidence that in most cases, except for the classification of texts from the WBLOG, the computed p-value is lower than the significance level $\alpha = 0.05$. As such, we are able to reject the null hypothesis and accept, with some confidence, the alternative hypothesis that values obtained by C-GMVLA for three views are shifted to the right of the values obtained by Co-training for two views.

	MPQA	RIMDB	CHES	WBLOG
p-values	$< 10^{-4}$	0.03	$< 10^{-4}$	0.3

Table 10: Wilcoxon rank-sum test for CHES.

Finally, to better understand the behaviour of the C-GMVLA algorithm, we illustrate in Figure 2 the different changes in classifiers due to changes in precision on the validation data set to choose the best one to classify unseen examples from the target domain. For the objective case, the best precision results at the beginning of the learning process are given by the Affective words view and then the best classifier is always the one based on unigrams. In this case, the level of abstraction view does not play any role. On the contrary, for the subjective part, the level of abstraction view guides the learning process until the unigram overtakes its precision levels at the sixth iteration. In this case, the Affective words view is useless for the subjective learning process. With these results, we clearly understand that multi-view learning for more than two views may provide better decisions when adding new examples to the labelled data and open new research trends.

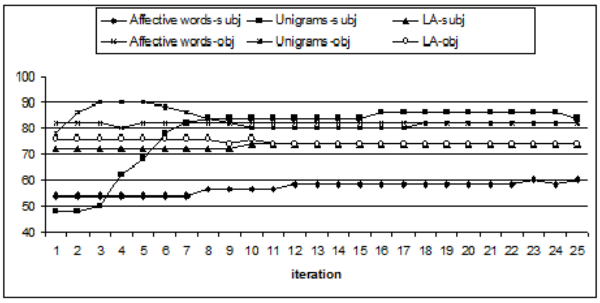

Figure 2: View precisions using C-GMVLA.

8. CONCLUSIONS AND FUTURE WORKS

In this paper, we proposed to use multi-view learning for cross domain subjectivity classification. We presented different experiments based on high-level and low-level features and showed that using more than two views can lead to improved results using the Class-Guided Multi-View Learning Algorithm with 91.3% accuracy. Nevertheless, some questions remain open. The main one is how to choose the methodology, which would find the most relevant combination of views to improve classification on a theoretical background. We are already working on that issue. Moreover, the comparison with the SAR algorithm is not fair as only low-level features can be used in the current framework. So, we aim to adapt the SAR algorithm to more than two views and different types of features. We already theoretically proved that the SAR algorithm can be adapted to 3-views, showing that the agreement function is proportional to the product of each probabilistic classifier. One other important idea is to combine both the SAR and the C-GMVLA algorithms into just one framework and as such benefit from both approaches. Finally, another idea is to use newly developed resources such as SenticNet 2 [6] as well as applying new ideas proposed in Sentic Computing [7].

9. 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] D. Bollegala, D. Weir, and J. Carroll. Using multiple sources to construct a sentiment sensitive thesaurus for cross-domain sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL/HLT 2011)*, pages 132–141, 2011.

[6] E. Cambria, C. Havasi, and A. Hussain. Senticnet 2: A semantic and affective resource for opinion mining and sentiment analysis. In *Proceedings of the 25th International FLAIRS Conference (FLAIRS 2012)*, 2012.

[7] E. Cambria and A. Hussain. *Sentic Computing: Techniques, Tools and Applications*, volume 2 of *Briefs in Cognitive Computation*. Springer, 2012.

[8] 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.

[9] 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.

[10] 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.

[11] X. Glorot, A. Bordes, and Y. Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th International Conference on Machine Learning (ICML 2011)*, pages 513–520, 2011.

[12] V. Hatzivassiloglou and K. McKeown. Predicting the semantic orientation of adjectives. In *Proceedings of the 8th Conference on European Chapter of the Association for Computational Linguistics (EACL 1997)*, pages 174–181, 1997.

[13] V. Hatzivassiloglou and J. Wiebe. Effects of adjective orientation and gradability on sentence subjectivity. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*, pages 299–305, 2000.

[14] Y. He, C. Lin, and H. Alani. Automatically extracting polarity-bearing topics for cross-domain sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011)*, pages 123—-131, 2011.

[15] T. Joachims. *Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms.* Kluwer Academic Publishers, 2002.

[16] T. Kailath. The divergence and bhattacharyya distance measures in signal selection. In *IEEE Transactions on Communications*, volume 15, page 52–60, 1967.

[17] B. Levin. *English Verb Classes and Alternations.* University of Chicago Press, 1993.

[18] H. Liu. Montylingua: An end-to-end natural language processor with common sense, 2004.

[19] G. A. Miller. Wordnet: an on-line lexical database. *International Journal of Lexicography*, 3(4), 1990.

[20] C. Osgood, G. Suci, and P. Tannebaum. *The Measurement of Meaning.* University of Illinois Press, 1971.

[21] S. Pan, X. Ni, J.-T. Sun, Q. Yang, and Z. Chen. Cross-domain sentiment classification via spectral feature alignment. In *Proceeding of the 19th International World Wide Web Conference (WWW 2010)*, pages 751–760, 2010.

[22] 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.

[23] 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 (EMNLP 2002)*, pages 79–86, 2002.

[24] L. Qiu, W. Zhang, C. Hu, and K. Zhao. Selc: a self-supervised model for sentiment classification. In *Proceeding of the 18th ACM Conference on Information and Knowledge Management (CIKM 2009)*, pages 929–936, 2009.

[25] J. Rocchio. Relevance feedback in information retrieval. In *The SMART Retrieval System: Experiments in Automatic Document Processing*, chapter 14, pages 313–323. Prentice-Hall, 1971.

[26] C. Strapparava and A. Valitutti. Wordnet-affect: An affective extension of wordnet. In *Proceedings of the 4th Language Resources and Evaluation International Conference (LREC 2004)*, pages 1083–1086, 2004.

[27] 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.

[28] J. Wiebe, T. Wilson, and C. Cardie. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2-3), 2005.

[29] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Joint Conference on Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, 2005.

[30] Q. Wu, S. Tan, X. Cheng, and M. Duan. Miea: a mutual iterative enhancement approach for cross-domain sentiment classification. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, pages 1327–1335, 2010.