Guided Self Training for Sentiment Classification

Brett Drury
LIAAD-INESC
Portugal
brett.drury@gmail.com

Luís Torgo
Fac. Sciences
LIAAD-INESC, Portugal
ltorgo@inescporto.pt

J.J Almedia
Dept. of Engineering
University of Minho, Portugal
jj@di.uminho.pt

Abstract

The application of machine learning techniques to classify text documents into sentiment categories has become an increasingly popular area of research. These techniques rely upon the availability of labelled data, but in certain circumstances the availability of pre-classified documents may be limited. Limited labelled data can impact the performance of the model induced from it. There are a number of strategies which can compensate for the lack of labelled data, however these techniques may be suboptimal if the initial labelled data selection does not contain a sufficient cross section of the total document collection. This paper proposes a variant of self-training as a strategy to this problem. The proposed technique uses a high precision classifier (linguistic rules) to influence the selection of training candidates which are labelled by the base learner in an iterative self-training process. The linguistic knowledge encoded in the high precision classifier corrects high-confidence errors made by the base classifier in a preprocessing step. This step is followed by a standard self training cycle. The technique was evaluated in three domains: user generated reviews for (1) airline meals, (2) university professors and (3) music against: (1) constrained learning strategies (voting and veto), (2) induction and (3) standard self-training. The evaluation measure was by estimated F-Measure. The results demonstrate clear advantage for the proposed method for classifying text documents into sentiment categories in domains where there is limited amounts of training data.

1 Introduction

The application of machine learning techniques to classify text into sentiment categories has become an increasingly popular area of research. Models induced from data can be very accurate(Halevy et al., 2009), but a learner may require a significant amount of data to induce a model which can accurately classify a text collection. Large volumes of labelled documents may not be readily available or may be expensive to obtain. Models induced from small volumes of labelled data may be suboptimal because the pre-classified data may not contain a sufficient cross-section of the document collection. The field of Semi-Supervised Learning (SSL) offers a number of possible strategies to compensate for the lack of labelled data. These techniques may not be effective if the model induced from the initial set of labelled data is biased or ineffective because these strategies can exacerbate the weaknesses in the initial model. This paper describes a SSL strategy that is a variant of Self-Training (ST). ST is an iterative process that obtains models with increasingly larger samples of labelled data. At each iteration the current model is used to classify the unlabelled data. The observations which the model has a high confidence in the classification are added to the next training sample with the classification of the model as the label.

The evaluation of the effectiveness of our proposal involved the experimental comparison with the following types of methods: (1) constrained, (2) inductive and (3) standard self- training. Training data was randomly selected from the total document collection and ranged from 1% of the total collection to 5%. The F-Measure was estimated by testing the model against the total document collection with the training data removed. The experiments were run 20 times for each training intervals with two separate learners: Naive Bayes and

Robust Unsupervised and Semi-Supervised Methods in Natural Language Processing, pages 9–16,
Hissar, Bulgaria, 15 September 2011.

Language Models. The domains which were evaluated were:(1) airline meals, (2) university teachers and (3) music reviews. The results demonstrate clear advantage for the proposed method for classifying text documents in domains where the models induced from the training data were weak.

1.1 Related Work

There are a number of approaches which use words (Hatzivassiloglou and McKeown, 1997)(Riloff and Weibe, 2003), phrases (Liu, 2007) and grammars (Drury and Almeida, 2011) to classify documents into sentiment categories. Linguistic rules may not be sufficient in domains which have non-standard linguistic features and lexicons. Another approach is to use labelled samples of the domain as training data to construct a classifier. Labelled data can be expensive to obtain. Semi-supervised learning can assist by adding labels to unlabelled data and using them as training data. A sub-field of semi-supervised learning uses constraints to limit the documents selected (Abney, 2007a). For example: co-training uses different views of the same data to train individual classifiers and restraining the documents selected to the documents which are labelled equally by the separate classifiers. The notion of "hard" constraints has been extended to with the idea of "soft" constraints (Druck et al., 2008). Druck (Druck et al., 2008) provides an example of using the Noun, "puck" to identify hockey documents. This type of soft constraint may not be successful with sentiment classification because separate classes can share features because a sentiment word can be negated. For example: the Noun, "recession" could be associated with a negative class, but with the addition of the word "v-shaped" transforms "recession" to a positive feature because the phrase "v-shaped recession" is positive, consequently the addition of the feature "recession" for the negative class would be an error. Chang (wei Chang et al., 2007) proposed the use of constraints to label a pool of unlabelled data and then use that pool of newly labelled data to update the model. Chang's approach would be insufficient for sentiment classification because of shared features where any individual unigram constraints could be negated and the pool of sentiment indicators are very large and that constraining the learner may produce a biased learner.

2 Proposed Strategy

The proposed ST variant - Guided Self-Training (GST) - differs from standard ST in the selection of the examples to add in each iteration. The variation is the use of a high precision classifier (linguistic rules) to select a small number of high confidence candidates ("the high confidence pool"). These rules are used to test the learner against the high confidence pool, and if the learner makes a high confidence erroneous classification of a member of this pool then the member is added to the correct class by the linguistic rules.

This training data is supplemented with extra data which the learner selects with high confidence from both non-members and members of the high confidence pool. The assumption of this proposed method is that the correction of erroneous high confidence classifications improves the performance of a learner and that the amount of improvement is directly related to the number of corrections. The learner is not explicitly constrained and is allowed to learn features from documents which are not in the high confidence pool, but the learner is "guided" to make correct selections when it makes serious errors.

2.1 Motivation

The principal motivation for this work was to identify a strategy which could construct a robust model which could classify documents into sentiment categories. Documents which are used for sentiment classification are often linguistically complex because they can contain: 1. multi word expressions which have semantic idiomaticity and 2. non standard spelling and grammar. These types of domains are difficult to classify because of the aforementioned features and because of the large volume of available documents to classify, for example Twitter claims that there are 50 millions tweets posted in a day[1]. It is not feasible to manually label or construct rules to label a significant number of these tweets. Learners which are constructed from a small subset of data are likely to be weak and traditional SSL techniques may not be suitable.

2.2 Problem Definition

GST is designed to improve the performance of a classifier in domains with the following characteristics:

[1] http://goo.gl/qXldd

Method	Avg. Precision
Method 1	57% ($\pm$3)
Method 2	75% ($\pm$3)

Table 1: Precision of Classifiers Induced from Rule Selected Data

- Limited amount of labelled data
- Labelling of large amounts of data is not feasible
- External resources such as general sentiment dictionaries (Esuli and Sebastiani, 2006) does not aid sentiment classification

2.3 Selection of high precision rule classifier

There are a number of methodologies to create a rule classifier. The rule classifier for GST must have a high precision and therefore recall was a secondary consideration. Two methodologies were considered: one which considered a sequence of POS tags to create bigrams (method 1) and the other which used manually selected features from training data and expanded them with Wordnet (Fellbaum, 1998) (method 2). These methodologies are described in detail by Liu (Liu, 2007). The competing methods selected and labelled data from reviews for airline food. Method 1 labels a document as according to the average opinion orientation (Liu, 2007). Method 2 labels a document as positive or negative if it has at least a difference of three unigrams from a given class. The difference of three unigrams produces an accurate classifier, but at the expense of recall (Riloff and Weibe, 2003).

The selection of the high precision classifier was by precision score of Language Models induced from the data selected by each competing methodology. A mean average precision score was calculated from a 2 X 5 cross validation process. The results are described in Table 1.

In this context, the GST method will use Method 2[2] to construct a dictionary for the high precision classifier because it has a higher precision than method 1. The rule classifier for GST will classify documents in the same manner as the rule selection test, i.e. review must have at least a difference of three unigrams from a given class. The proposed GST method is not dependent on this rule construction methodology, but any alternative rule classifier must have high precision which is normally at the cost of low recall.

2.4 GST Algorithm

The proposed GST method is described in Algorithm 1. GST takes two main inputs: the labelled (LD) and unlabelled (UD) data sets. The outer loop (lines 3-26) represent the typical self-training iterations. The uniqueness of the proposal are the following:

- Documents classified by the base learner with a high confidence which are contrary to the high precision classification (the pool of high confidence candidates) are assigned to the high precision classification. These documents are assigned to the labelled data for training in the next iteration.
- The high precision classifier can abdicate (i.e no decision) and therefore high confidence candidates can be selected by the base learner with out the explicit agreement of the high precision classifier.

A model is induced from the selected data. At each iteration, weaknesses in the model are corrected, but the document selection is not constrained to the pool of high confidence candidates (high precision classifier classifications), and consequently the learner reaches its optimum performance with less training data than competing methods.

3 Experimental Evaluation

Three domains were chosen for the evaluation of the proposed technique: (1) user generated reviews of airline meals (airlinemeals.net, 2010), (2) user generated reviews of university lecturers (ratemyprofessors.com, 2010) and (3) user generated reviews of music concerts and records (reviewcentre.com, 2010) [11][3]. The domains demonstrated the following linguistic characteristics: (1) invented words, (2) slang, (3) profanity, (4) non standard spelling and grammar, (5) multi-word expressions (MWE) and (6) non standard punctuation.

[2]Method 2 recorded an average precision of 95% and recall of 20% when the rules directly classified candidate data and were allowed to abdicate.

[3]Data and dictionaries can be found at http://goo.gl/IHL6V

Algorithm 1 Description of GST Candidate Selection Cycle

procedure GST($LD, UD, sThr, Rules, Learner$)
▷ LD and UD - The collections of labelled an unlabelled documents, respectively; $sThr$ - The minimum classification confidence for a document to be considered for addition to the labelled training set; $Rules$ - A series of linguistic rules which return a classification for a document; $Learner$ - the classification algorithm that is to be self-trained. CD - a container for a corrected documents - i.e. errors made by the base classifier AD A container for documents where the base and high precision classifier don't disagree TD - a container for documents in CD which are not selected for training

$Model \leftarrow Learner(LD)$ ▷ Learn a classifier
repeat
$lClass \leftarrow Model.classify(UD)$
$rClass \leftarrow Rules.classify(UD)$
$CD \leftarrow \{\}$
$AD \leftarrow \{\}$
▷ Check agreement between Learner and Rules
for all $d \in UD$ **do**
if $lClass.confidence[d] \geq sThr$ **then**
$UD \leftarrow UD \setminus d$
if $rClass[d] \neq NULL$ **and** $rClass[d] \neq lClass[d]$ **then**
$CD \leftarrow CD \cup \{< d, rClass[d] >\}$
else
$AD \leftarrow AD \cup \{< d, lClass[d] >\}$
end if
end if
end for
$count \leftarrow Count(CD)$
if $count == 0$ **then**
$count \leftarrow Count(AD)$
end if
$TD \leftarrow ReturnRandomDocs(AD, count)$
$UD \leftarrow UD \cup (AD \setminus TD)$
$LD \leftarrow LD \cup CD \cup TD$
$Model \leftarrow Learner(LD)$ ▷ Get a new model
until $terminationCriterion$
return $Model$
end procedure

3.1 Experimental Setup

Each document contained: the text and a form of rating. The rating was taken as an indication of the polarity of the review. The criteria for class assignment is described in Table 2. Documents not satisfying the criteria for class assignment were removed from our experiments.

These resulting labelled data sets were used to compare:

- Two separate base learners (Naive Bayes and Language Models)

Domain	Positive Category	Negative Category
Airline Meals	4 -5 Stars	1-2 Stars
Teacher Reviews	Good Quality	Poor Quality
Music Reviews	4-5 Stars	1-2 Stars

Table 2: Polarity Criteria

- Alternative strategies

The evaluation was by means of an estimated F-Measure. The experiments used increasing larger random selection of documents as training data. The smallest selection of data was 1% of the total and the largest 5%. The increments were in steps of 1%, for example the second iteration of the experiment was 2%, the third 3% etc. At each iteration the experiment was repeated 20 times, for example the 1st iteration there would be 20 random samples of 1% and 20 estimations of F-Measure. An overview of the process is the following: 1. randomly select training data (the LD set in Algorithm 1) and 2. "artificially unlabel" the remaining documents to create the UD.

The experiments were repeated using Language Models and Naive Bayes Classifier as the baseline classifiers within the GST algorithm.

We have compared our proposed method against three alternative strategies:

(1) inductive, (2) self-training and (3) constrained learning.

- Inductive: An inductive strategy induces a classification model using only the labelled data (Abney, 2007b).

- Self-Training: An iterative process where at each step a model is induced from the current labelled data and it is used to classify the unlabelled data set. The model assigns a "confidence measure" to each classification. If the classification confidence measure is greater than a predefined threshold then the respective unlabelled cases are added to the new iteration training data with the classifier assigned label. At the end of the cycle the learner is trained on the "new labelled data set". This cycle continues until a stopping condition is met (Abney, 2007b). To ensure an equitable comparison the stopping condition for both self-training and GST was 5 iterations.

- Constrained Learning: The alternate constrained learning strategies were Voting and Veto.
  - Voting strategy: Selects documents if both the classifiers agree on the classification of the document
  - Veto strategy: The base learner selects the data, but high precision classifier adds the label, consequently the high precision classifier vetoes a dissenting leaner classification. The high precision classifier is not allowed to abdicate.

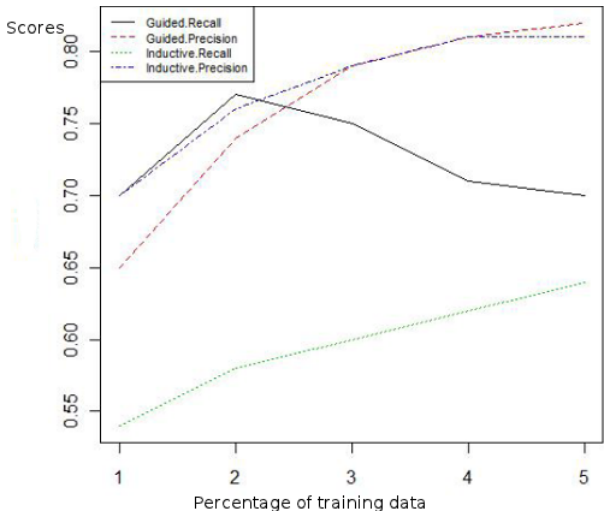

Figure 1: Language Models: Comparative Recall and Precision for Teacher Domain

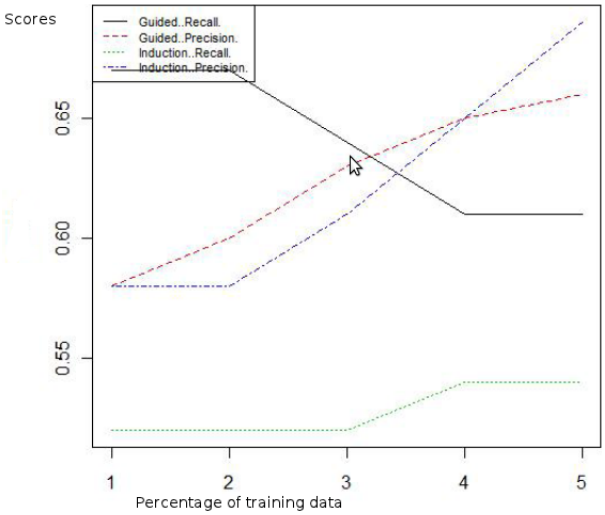

Figure 2: Naive Bayes: Comparative Recall and Precision for Airline Meals Domain

4 Experimental Results

The *Airline Food Domain* results are presented in Table 3. The results demonstrate a clear advantage for the proposed strategy for both classifiers. The results demonstrate a significant gain in F-Measure at the 2% of domain for training for both classifiers. The gain in F-Measure halts at the 3% of domain for training. The two inductive strategies gain F-Measure as training data increases.

The *Teachers Domain* results are presented in Table 4. The results demonstrate a clear advantage

		% of Data for Training				
		1	2	3	4	5
Algorithm	Classifier	F-Measure	F-Measure	F-Measure	F-Measure	F-Measure
Fully Supervised	Naive Bayes	0.91				
Fully Supervised	Language Models	0.98				
GST	Naive Bayes	0.52 ±0.05	0.61 ±0.01	0.63 ±0.01	0.63 ±0.01	0.63 ±0.01
GST	Language Models	0.49 ±0.04	0.60 ±0.02	0.64 ±0.01	0.64 ±0.01	0.63 ±0.02
Voting	Naive Bayes	0.48 ±0.00	0.49 ±0.00	0.50 ±0.01	0.51 ±0.01	0.51 ±0.01
Voting	Language Models	0.48 ±0.00	0.49 ±0.00	0.49 ±0.00	0.50 ±0.00	0.51 ±0.00
Inductive (LD)	Naive Bayes	0.51 ±0.01	0.51 ±0.01	0.52 ±0.01	0.54 ±0.01	0.55 ±0.01
Inductive (LD)	Language Models	0.49 ±0.02	0.50 ±0.01	0.51 ±0.01	0.52 ±0.01	0.53 ±0.01
Inductive (LD+RC)	Naive Bayes	0.54 ±0.00	0.55 ±0.00	0.56 ±0.00	0.56 ±0.00	0.57 ±0.00
Inductive (LD+RC)	Language Models	0.53 ±0.00	0.54 ±0.00	0.55 ±0.00	0.55 ±0.00	0.56 ±0.00
Self-Training (LD)	Naive Bayes	0.50 ±0.01	0.50 ±0.01	0.51 ±0.01	0.51 ±0.01	0.52 ±0.01
Self-Training (LD)	Language Models	0.48 ±0.01	0.49 ±0.00	0.50 ±0.00	0.50 ±0.01	0.51 ±0.00
Veto	Naive Bayes	0.54 ±0.00	0.55 ±0.00	0.56 ±0.00	0.49 ±0.00	0.49 ±0.00
Veto	Language Models	0.53 ±0.00	0.54 ±0.00	0.55 ±0.00	0.55 ±0.00	0.56 ±0.00

Table 3: Airline Meals Experimental Results

for the proposed strategy. In common with the airline food domain the Guided Self-Training(GST) shows a large gain in F-Measure at 2% of domain for training. The gain in F-Measure is more pronounced for language models. GST demonstrates a reduction in F-Measure with further increases in training data. The reduction in F-Measure is within the mean standard deviation. The inductive strategists in common with the airline food domain gains F-Measure with increases in training data. The self-training strategy gains in F-Measure increase with training data, but at a faster rate than the inductive strategies. The voting schemes also demonstrate a gain in F-Measure, but at a lower rate than the inductive and self-training strategies.

The *Music Review Domain* results are presented in Table 4. The results demonstrates that the proposed strategy does not show any distinct advantage over the competing strategies. The models induced from the labelled data seem robust and the various SSL strategies fail to improve this strategy.

4.1 Discussion of Results

Strategies which have access to rule selected data frequently have a higher precision measure, but this improvement is frequently at the cost of lower recall. For example the mean average recall and precision for the voting strategy in the Airline Food domain was 0.5 and 0.7, where as the inductive strategy yielded recall and precision of: 0.51 and 0.62. A possible explanation for this phenomenon is that fact that the high precision classifier may only classify a very specific sample of documents. The addition of these documents labelled by the high precision classifier to the initial data set of the models we could be biasing the classifier towards learning very specific rules, which may negatively impact on recall, but may boost precision. The GST method does not suffer from a decrease in recall. A possible explanation could be the high precision classifier is being used with a different purpose within GST when compared to the (LD+RC) learners. In GST high precision classifier are used to supervise the classifications of a standard base learner with the goal of avoiding very obvious mistakes. In the (LD+RC) learners the rules are used to add more labelled data to the training set available to the learners. These are two different uses of the high precision classifier and our experiments clearly provide evidence towards the advantage of our proposal. In effect, GST improvement in precision is not offset by a drop in recall.

The graphs illustrated in Figure 2 and Figure 1 provide a comparative analysis of the precision and recall for the inductive and proposed strategy in the airline and teacher domains respectively. These graphs provide some evidence for the assertion that the F-Measure gains are at not at the expense of a drop in recall because until 2% domain training data there are gains in precision and no drop in recall. The airline domain demonstrates a gain in recall. The recall drops from 2% onwards, however recall is always significantly higher than the recall for the inductive strategy. The GST strategy continues to gain precision with increases in training data.

4.2 Discussion of Methodology

The assumption of the GST methodology is that correcting high confidence erroneous classifications and including the documents as training

	% of Data for Training					
		1	2	3	4	5
Algorithm	Classifier	F-Measure	F-Measure	F-Measure	F-Measure	F-Measure
Fully Supervised	Naive Bayes	0.96				
Fully Supervised	Language Models	0.99				
GST	Naive Bayes	0.67 ±0.04	0.71 ±0.02	0.67 ±0.03	0.66 ±0.02	0.65 ±0.02
GST	Language Models	0.58 ±0.01	0.75 ±0.01	0.76 ±0.01	0.74 ±0.02	0.73 ±0.02
Voting	Naive Bayes	0.47 ±0.01	0.48 ±0.01	0.51 ±0.01	0.52 ±0.01	0.54 ±0.01
Voting	Language Models	0.45 ±0.01	0.48 ±0.01	0.49 ±0.01	0.51 ±0.01	0.53 ±0.01
Inductive (LD)	Naive Bayes	0.56 ±0.03	0.60 ±0.02	0.63 ±0.03	0.65 ±0.02	0.66 ±0.02
Inductive (LD)	Language Models	0.52 ±0.02	0.59 ±0.03	0.61 ±0.02	0.64 ±0.02	0.66 ±0.02
Inductive (LD+RC)	Naive Bayes	0.53 ±0.00	0.54 ±0.00	0.54 ±0.05	0.57 ±0.00	0.58 ±0.00
Inductive (LD+RC)	Language Models	0.52 ±0.00	0.53 ±0.00	0.55 ±0.00	0.56 ±0.00	0.57 ±0.00
Self-Training (LD)	Naive Bayes	0.53 ±0.03	0.56 ±0.02	0.60 ±0.02	0.62 ±0.03	0.64 ±0.02
Self-Training (LD)	Language Models	0.49 ±0.02	0.55 ±0.03	0.57 ±0.02	0.60 ±0.02	0.62 ±0.02
Veto	Naive Bayes	0.52 ±0.00	0.54 ±0.00	0.55 ±0.00	0.57 ±0.00	0.58 ±0.00
Veto	Language Models	0.52 ±0.00	0.53 ±0.00	0.55 ±0.00	0.56 ±0.00	0.57 ±0.00

Table 4: Teacher Review Experimental Results

	% of Data for Training					
		1	2	3	4	5
Algorithm	Classifier	F-Measure	F-Measure	F-Measure	F-Measure	F-Measure
Fully Supervised	Naive Bayes	0.96				
Fully Supervised	Language Models	0.99				
GST	Naive Bayes	0.51 ±0.01	0.46 ±0.02	0.48 ±0.02	0.49 ±0.02	0.50 ±0.03
GST	Language Models	0.54 ±0.01	0.55 ±0.01	0.49 ±0.01	0.49 ±0.02	0.48 ±0.02
Voting	Naive Bayes	0.43 ±0.00	0.44 ±0.00	0.45 ±0.01	0.46 ±0.01	0.47 ±0.01
Voting	Language Models	0.43 ±0.00	0.45 ±0.01	0.45 ±0.01	0.46 ±0.01	0.47 ±0.01
Inductive (LD)	Naive Bayes	0.57 ±0.09	0.64 ±0.07	0.61 ±0.07	0.66 ±0.07	0.65 ±0.06
Inductive (LD)	Language Models	0.54 ±0.09	0.59 ±0.11	0.60 ±0.07	0.65 ±0.08	0.67 ±0.06
Inductive (LD+RC)	Naive Bayes	0.45 ±0.00	0.45 ±0.00	0.46 ±0.01	0.47 ±0.01	0.48 ±0.01
Inductive (LD+RC)	Language Models	0.45 ±0.01	0.46 ±0.00	0.46 ±0.00	0.47 ±0.00	0.48 ±0.01
Self-Training (LD)	Naive Bayes	0.58 ±0.01	0.64 ±0.07	0.61 ±0.07	0.66 ±0.08	0.65 ±0.06
Self-Training (LD)	Language Models	0.54 ±0.09	0.58 ±0.12	0.60 ±0.08	0.65 ±0.09	0.66 ±0.07
Veto	Naive Bayes	0.45 ±0.00	0.45 ±0.00	0.46 ±0.01	0.47 ±0.01	0.48 ±0.01
Veto	Language Models	0.45 ±0.00	0.46 ±0.01	0.46 ±0.00	0.47 ±0.00	0.48 ±0.01

Table 5: Music Reviews Experimental Results

data will improve the performance of the induced model. An experiment was conducted where the number of documents corrected was recorded per training cycle. The experiment was conducted for the 1% of domain for training. The results are described in Figure 3. The two domains in which GST gained the highest F-Measure there is a high level of corrections in the first training cycle. The remaining cycles show a small number of corrections. The music domain demonstrates a small number of corrections and may account for the relatively poor performance of GST in this domain.

The second assumption of this methodology is that the classifier which corrects the erroneous classifications must be accurate and that classifiers with lower precision will effect the performance of the GST strategy. A further experiment was conducted with a high precision classifier which was constructed with a lower precision methodology (method 1). The experiment was conducted on the airline meals domain data. The results are presented in Table 6. The results were the worst of all of the strategies tested. The lower precision classifier "modified" correct high confidence classifications made by the base learner rather than the high confidence erroneous classifications. The "erroneous corrections" inhibited the base learner and induced a weak model. Although the aforementioned methodology had an inferior precision than the classifier used for the proposed strategy its overall performance was slightly better. Table 6 shows the inductive(RD+LC) strategy which used rule selected data gained a slightly higher F-Measure than for the inductive(RD+LC) strategy in the main experiments (Table 3).

5 Conclusion

This paper describes a new semi-supervised classification method for sentiment classification (GST) designed with the goal of handling document clas-

	% of Data for Training					
		1	2	3	4	5
Algorithm	Classifier	F-Measure	F-Measure	F-Measure	F-Measure	F-Measure
GST	Language Models	0.13 ±0.00	0.15 ±0.00	0.17 ±0.00	0.21 ±0.00	0.25 ±0.00
Inductive (RD+LC)	Language Models	0.58 ±0.00	0.58 ±0.00	0.59 ±0.00	0.59 ±0.00	0.59 ±0.00

Table 6: GST Strategy with lower precision classifier

sification tasks where there are limited labelled documents. The proposed technique can perform well in circumstances where more mature strategies may perform poorly. The characteristics of the domains where it is thought that this strategy will offer a clear advantage are the following: (1) model induced from labelled data makes obvious mistakes, (2) adding more data (either manually or by rules) does not improve performance, and (3) it is possible to construct a high precision rule based classifier. GST uses linguistic information encoded into a high precision classifier. This information is not added on mass where the learner may be biased towards information captured by the high precision classifier, but it is added in areas where the learner is weak. GST also selects a larger variety of documents than the high precision classifier because the base learner self-selects high confidence candidates from the unlabelled data. The constant testing of the learner prevents drift which may occur in classical self-training. The proposed technique provides a viable alternative to current semi-supervised classification strategies, as our experimental results demonstrate.

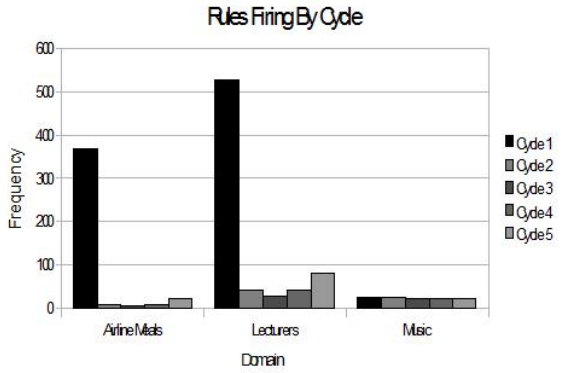

Figure 3: No. Corrected Documents Per Training Cycle.

References

S. Abney. 2007a. *Semisupervised Learning for Computational Linguistics*. Chapman & Hall/CRC. Chapter 9: Agreement Constraints.

S. Abney. 2007b. *Semisupervised Learning for Computational Linguistics*. Chapman & Hall/CRC. Chapter two: Self Training and Co-Training.

airlinemeals.net. 2010. Airlinemeals. `http://www.airlinemeals.net/`, consulted in 2010.

Gregory Druck, Gideon S. Mann, and Andrew Mccallum. 2008. Learning from labeled features using generalized expectation criteria. In *Research and Development in Information Retrieval*, pages 595–602.

Brett Drury and José João Almeida. 2011. Identification of fine grained feature based event and sentiment phrases from business news stories. In *WIMS*, page 27.

A. Esuli and F. Sebastiani. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. In *In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC 06*, pages 417 – 422.

C. Fellbaum. 1998. *WordNet: An Electronical Lexical Database*. The MIT Press, Cambridge, MA.

Alon Halevy, Peter Norvig, and Fernando Pereira. 2009. The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 2:8–12.

Vasileios Hatzivassiloglou and Kathleen R. McKeown. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 174–181.

Bing Liu. 2007. *Web Data Mining: chapter(Opinion Mining)*. Springer.

ratemyprofessors.com. 2010. Ratemyprofessors. `http://www.ratemyprofessors.com/`, consulted in 2010.

reviewcentre.com. 2010. Reviewcentre. `http://www.reviewcentre.com/`, consulted in 2010.

E. Riloff and J. Weibe. 2003. Learning extraction patterns for subjective expressions. In *In Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 105–112.

Ming wei Chang, Lev Ratinov, and Dan Roth. 2007. Guiding semi-supervision with constraint-driven learning. In *In Proc. of the Annual Meeting of the ACL*.