Reserved Self-training: A Semi-supervised Sentiment Classification Method for Chinese Microblogs

Zhiguang Liu Xishuang Dong Yi Guan* Jinfeng Yang
School of Computer Science and Technology
Harbin Institute of Technology, Harbin, China
{triggerliu, dongxishuang, yangjinfeng2010}@gmail.com
guanyi@hit.edu.cn

Abstract

The imbalanced sentiment distribution of microblogs induces bad performance of binary classifiers on the minority class. To address this problem, we present a semi-supervised method for sentiment classification of Chinese microblogs. This method is similar to self-training, except that, a set of labeled samples is reserved for a confidence scores computing process through which samples that are less than a predefined confidence score threshold are incorporated into training set for retraining. By doing this, the classifier is able to boost the performance on the minority class samples. Experiments on the NLP&CC2012 Chinese microblog evaluation data set demonstrated that reserved self-training outperforms the best run by 2.06% macro-averaged and 2.30% micro-averaged F-measure, respectively.

1 Introduction

Sentiment classification aims to label peoples opinions as different categories such as positive and negative from a given piece of text (Pang et al., 2002). Currently, related research on traditional online media, such as blogs, forums, and online reviews, has made great progress (Banerjee and Agarwal, 2012; Liu et al., 2005). However, sentiment classification of microblogs is hard to process due to some unique characteristics of microblogs, for example, short length of update messages and language variations. Moreover, topic based microblogs are related with peoples daily lives and people are more likely to post some negative messages to show their unsatisfactoriness, which may partially result in imbalanced sentiment class distributions. For example, the number of negative tweets is far more than that of positive in some topics, which is different from the previous work on sentiment classification that assumes the balance between positive and negative samples (Chawla et al., 2002; Yen and Lee, 2009).

While supervised techniques have been widely used in sentiment classification (Pang et al., 2002), the main problem that supervised methods suffered is that they rely on labeled data solely. Semi-supervised methods, which make use of both labeled and unlabeled data, are ideal for sentiment classification, since the cost of labeling data is high whereas unlabeled data are often readily available or easily obtained (Ortigosa-Hernández et al., 2012). However, there are some drawbacks of semi-supervised approaches such as most of the work assume that the positive and negative samples in both labeled and unlabeled data set are balanced, otherwise models often bias towards the majority class (Chawla et al., 2002; Yen and Lee, 2009). In addition, most existing studies on imbalanced classification focus on supervised learning methods, with few on semi-supervised approaches (Li et al., 2011).

In this study, we propose a reserved self-training method for binary sentiment classification inspired by active learning strategies (Ryan, 2011). Active learning systems interact with domain experts who are responsible for annotating unlabeled samples, and aim to achieve better performance with less training data (Wu and Ostendorf, 2013). The key to active learning is to find an appropriate query strategy such as the classifier poses queries to decide which samples are most informative. We randomly reserved a portion of labeled samples before training. Reserved self-training is the process of simulating active learning that repeatedly queries our reserved samples and then incorporates the labeled samples about which the classifier is least certain into training corpus for retraining, thus the retrained classifier is able to improve

*Corresponding author: guanyi@hit.edu.cn

International Joint Conference on Natural Language Processing, pages 455–462,
Nagoya, Japan, 14-18 October 2013.

the performance of classification on the minority class.

The remainder of this paper is organized as follows. The next section reviews some related work on semi-supervised sentiment classification as well as imbalanced classification briefly. We formally define the task in section 3. Section 4 presents our approach of reserved self-training algorithm for imbalanced sentiment classification. Section 5 provides experimental results on a data set of 20 topics. Finally, section 6 summarizes the work, draws some conclusions, and suggests related future work.

2 Related Work

Sentiment classification ranges from the document level, to the sentence and phrase level, and we concentrate on sentence level classification. Sentence level sentiment classification methods can be categorized into three types: supervised (Pang et al., 2002) unsupervised (Turney, 2002), and semi-supervised learning methods (Singh et al., 2008) among which semi-supervised approaches are more appropriate for sentiment classification of microblogs due to their capability of making use of both labeled and unlabeled data. Another related work is imbalanced classification stems from several unique characteristics possessed by microblogs (A detailed study can be found in section 3.2).

2.1 Related Semi-supervised Sentiment Classification Works

Semi-supervised learning approaches make good use of a small portion of labeled and a large amount of unlabeled data to build a better classifier. One of the bottlenecks in applying supervised learning is that it needs to label many samples by domain experts. To save the work of manual annotation, Riloff et al. (2003) introduced a bootstrapping method which was able to automatically label training samples. They started on a few seeds for training, subsequently, incorporated five highest scores unlabeled samples into training corpus to retrain the model iteratively. Chang et al. (2007) added some restrictions to self-training, making it possible to produce better feedback information in the learning process. For a given classification task, one of the problems of adopting co-training is that it assumes two conditionally independent feature sets could be extracted (Blum and Mitchell, 1998). Although further studies loosed this strong assumption (Balcan et al., 2004), two classifiers must be different enough to achieve complementation. Li et al. (2011) proposed a random subspace generation algorithm for co-training applied to imbalanced sentiment classification, but its corpus limited to English product reviews.

2.2 Related Imbalanced Classification Works

Imbalanced classification, as an appealing task, has been extensively studied in many research areas such as pattern recognition (Barandela et al., 2003) and data mining (Chawla et al., 2004). We pay special attention to resampling and cost-sensitive methods, since they are widely applied in imbalanced classification. Other methods such as induction technique and boosting (Weiss, 2004) are beyond the scope of this paper.

Resampling is a process in which the size of training samples is changed to modify the overall size and distribution of a corpus, among these methods downsampling and oversampling are two widely used resampling techniques. Downsampling (Barandela et al., 2003) takes a subset of majority classes samples whereas oversampling (Chawla et al., 2002) randomly repeats minority classes samples to keep balance between different classes. Downsampling needs shorter training time, at the expense of disregarding potentially useful samples. Oversampling increases the size of training data set that leads to a longer training time. Moreover, oversampling may cause over fitting due to minority class samples are randomly duplicated (Chawla et al., 2002; Drummond et al., 2003). In addition to the basic downsampling and oversampling techniques, there are some other sampling methods working in a more complicated fashion. SMOTE (Chawla et al., 2002) created some synthetic minority class examples and then performed a combination of oversampling and downsampling, which achieved better performance than only applying downsampling. Some other methods integrated different sampling strategies to obtain further improvement(Batista et al., 2004).

Cost-sensitive learning (Ling et al., 2004; Zadrozny et al., 2003) is another type of method used for dealing with imbalanced classification. Most cost-sensitive learning methods can be generally divided into two categories (Lee et al., 2012): transforming an existing cost-insensitive

classifier into an equivalent cost-sensitive via a wrapper approach, or taking the cost of misclassification into consideration when training a classifier by labeled samples.

3 Task Definition

We first give a formal definition of our task, and then analyze the unique characters of Chinese microblogs compared to traditional online media, such as forums and blogs.

3.1 The Task

Our study involves classifying opinions of Chinese microblogs as either positive or negative. We perform sentence level sentiment classification for a given message of microblogs. We first conduct some preprocessing such as word segmentation and noisy symbols filtering. Subsequently, features for the classifier are extracted from each message. Finally, reserved self-training is employed to predict unlabeled data. Although we restrict the scope of study on Chinese microblogs, the method proposed in this study can be straightly extended in support of other languages such as English.

Here is an instance of illustrating our task. For a message "#50个人生必去的胜地# 万里长城太棒了，将来一定去看看！ @李雷" (#The 50 places you must see# The Great Wall of China is amazing, I will visit it someday in future@ Lei Li, a Chinese name). The words between the # symbol refers to a relevant topic and the symbol @ means a mention or reply. This message is expected to be parsed into a triple: (Topic: The 50 places you must see), (Content: The Great Wall of China is amazing, I will visit it someday in future), (Polarity: Positive). Here, 'Topic' is a key word people interested; 'Content' refers to the content of a posted message; 'Polarity' denotes the predicted polarity produced by our model, and the possible values for it could be positive and negative.

3.2 Characteristics of Sentiment Classification of Chinese Microblogs

Compared with traditional media such as blog and product reviews, detecting sentiment from microblogs is much harder due to the following challenges posed by microblogs. First, different from English, each written Chinese sentence need to be split into a sequence of words, however, the frequent use of informal and irregular words in microblogs may hinder the accuracy of segmentation. Second, the short length of messages and language variation contribute to the data sparsity problem. Third, different from previous work which concentrated on specific domain such as digital product reviews, sentiment classification of microblogs involves multi-domain information, thus, the model trained on one domain may perform badly when shift to another one. Lastly, the dynamic updates feature of microblogs means that sentiment class distributions may vary over time, in which case we need to handle imbalanced sentiment classification.

The NLP&CC2012[1] evaluation data set consists of 20 topics collected from Tencent Microblog[2], involving multiple domains such as political, environmental, and health issues. Illustrated in Figure 1, it can be observed that all the classes of training corpus are biased. In particular, positive sentences account for the majority class in topic 3, 6, and 11, which is different from the other topics.

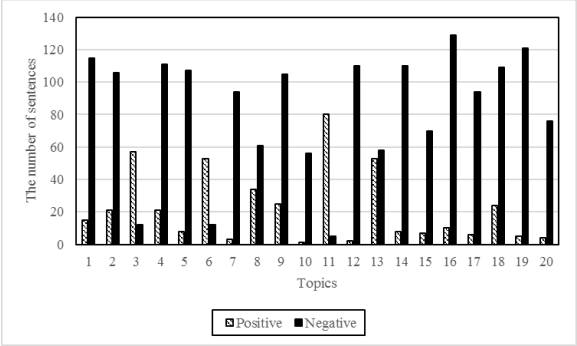

Figure 1: Class distributions of positive and negative samples in 20 topics

4 Reserved Self-Training for Imbalanced Classification

In this study, we incorporate a learning strategy into self-training, inspired by active learning, to tackle the imbalanced binary classification problems.

4.1 Self-training

Self-training is a common method of semi-supervised learning which makes use of both the labeled and unlabeled data as training corpus. As shown in Algorithm 1, Self-training is a wrapper algorithm that iteratively applies supervised

[1] http://tcci.ccf.org.cn/conference/2012
[2] http://t.qq.com

Algorithm 1 The self-training algorithm

Input:

Labeled data L

Unlabeled data U.

Procedure:

1. Apply supervised method to train a classifier C with L.
2. Make predictions on unlabeled data U with C.
3. Incorporate the most confidently predicted unlabeled data M in U along with each predicted label into $L = L \cup M$.
4. Loop for S iterations.

Output:

New labeled sample set L and classifier C.

method inside. It starts training on labeled data only, after each iteration, the most confidently predicted unlabeled samples would be incorporated as additional labeled data, decided by confidence scores calculation function. However, applying self-training to sentiment classification of Chinese microblogs in both subjectivity detection and sentiment classification performed not as well as expected, and the prediction results often bias towards the majority class. Comparing with fully supervised methods, the performance of self-training is even worse especially on the minority class(A detailed comparative study can be found in Section 5.2). It is not economical to revise the supervised classifier inside self-training, however, we may improve the data selection strategy to boost the performance of self-training on the minority class samples. Reserved self-training is such a technique that applies selection strategy in both labeled and unlabeled data during the learning process.

4.2 Reserved Self-training Classification Algorithm

In some cases, it seems unclearly what self-training is really doing, and which theory it corresponds to (Chapelle et al., 2006). Intuitively, it is almost definite to label high confidence samples, namely with little effect on the model. However, the discriminative ability of the model could be significantly improved if we try to label those samples about which the classifier is least certain. Similar to self-training, the idea behind reserved self-training is quite simple except that we first reserve a portion R of the training set L before training the initial classifier. As depicted in Algorithm 2, we apply the classifier to predict the unlabeled data U and the reserved data R, then we add those most confident unlabeled data and those least confident reserved data into training set T'. By adding training samples in this way, the classifier could increase the coverage of its decision space while not adding too many majority class samples. We use training set T' to train the model C iteratively until stopping criterion is met. Finally, assessing the performance of classifier C' on a labeled data set.

Algorithm 2 The algorithm of reserved self-training for imbalanced sentiment classification

Input:

Labeled data L consisting of positive examples P and negative examples N, where $|N| > |P|$.

Unlabeled data U that is also imbalanced.

Procedure:

Initialization:

1. Reserve a random portion R of L, and the remaining set $L' = L - R$ is used for training.
2. Loop for M iterations
3. Train the classifier C with L'.
4. Make predictions on unlabeled data U with C.
5. Predict the reserved portion of labeled data R by classifier C.
6. Incorporate the most confidently predicted unlabeled data in U along with each predicted label into L'.
7. Incorporate the least confidently predicted labeled samples in R into L'.

Output:

New labeled sample set L' and classifier C.

4.3 Labeled Data Selection

Generally, semi-supervised sentiment classification takes much less training data than supervised approaches, which forcing us to select the most effective samples from labeled data available. We resort to the principle of maximizing the diversity of samples in feature space to select seed. First, choose several samples as initial seed at random. Second, compute the centroid of the seed in feature space. Lastly, select those samples with least similarity to centroid of the seed done by cosine

similarity. By choosing seed in this manner, we aim to build a diversified data set to cover the feature space properly.

4.4 Confidence Scores Calculation

For binary classification, we employ probabilistic model to determine the confidence to which class a given sentence belongs, in that case the classifier queries the samples whose posterior probability of being positive or negative is nearest to pre-defined threshold. In this study, we employ MaxEnt and SVM as basic polarity classification. Normally, we could obtain the predicted label along with their confidence scores by MaxEnt. SVM adopt linear model to classify new examples, because of which we could use distances between samples and separating hyperplane to represent confidence scores (Pang and Lee, 2004). The output d_i of SVM is a signed distance (negative = negative orientation) from hyperplane, we convert d_i to non-negative by equation (1).

$$P_{neg}(s) = \begin{cases} 1 & d_i > 1 \\ (1+d_i)/2 & -1 \leq d_i \leq 1 \\ 0 & d_i < -1 \end{cases} \quad (1)$$

5 Experiments

This section details the experimental setup, including the corpora and lexicons we used, and the achieved results.

5.1 Experimental Setup

Benchmark Datasets: Our experiments are based on the Chinese Microblogs Sentiment Analysis Evaluation benchmark, China Computer Federation Conference on Natural Language Processing & Chinese Computer (NLP&CC2012). The evaluation is part of the NLP&CC2012, consisting 20 topics provided by Tencent Microblog , and there are 2207 subjective, 407 positive, and 1766 negative sentences.

Sentiment Lexicon: In our experiments, we integrate the following resources to construct a sentiment lexicon: (1) Sentiment lexicon provided by HowNet[3] which consists of 836 positive sentiment words and 1254 negative sentiment words; (2) NTU Sentiment Dictionary[4] from National Taiwan University. It contains 2,812 positive words and 8,276 negative words; (3) WI sentiment analysis lexicon[5] constructed by Harbin Institute of Technology which consists of 1,428 sentiment words with sentiment scores.

Feature selection As described in section 3.2, microblogging services is different from traditional media such as blog and product reviews. Special features should be explored according to the characteristics of microblogs, the main features we used can be found in table.1.

Table 1: The main features for polarity classification of opinion sentences

NO.	Feature description	Example
1	sentiment words	good, bad
2	strength of sentiment words	strength of pleasure, anger, sorrow, fear
3	rhetorical structure	question
4	emoticons	^_^
5	preposition	it, he, she
6	slang	给力(geli)[6]
7	repeated punctuation	!!!, ???
8	condition operator relates to a sentiment statement	despite, however, negative operator

5.2 Experimental Results

Supervised Learning for Imbalanced Sentiment Classification of Chinese Microblogs

In this section, we perform SVM and MaxEnt as our basic polarity classifier for sentiment classification of Chinese microblogs. Downsampling and oversampling are two widely used resampling technique for imbalanced classification, thus for thorough comparison, we apply SVM and MaxEnt model based on full training, downsampling, and oversampling method, depicted as follows.

1) Full-training: using the entire labeled corpora for training.

2) Downsampling: drop some of the majority class samples at random to obtain a balanced data set.

3) Oversampling: randomly duplicate the minority class samples to keep balance between the majority class and minority class.

[3] http://www.keenage.com

[4] http://nlg18.csie.ntu.edu.tw

[5] http://wi.hit.edu.cn

[6] "geili" is a Chinese word in English alphabet, which means something is cool, or cooperative.

Table 2: Performances of different methods for imbalanced sentiment classification

Approach		Evaluation metrics					
		Micro-average			Macro-average		
		Precision	Recall	F-Measure	Precision	Recall	F-Measure
SVM	full-training	0.8542	0.6364	0.7294	0.8581	0.6312	0.7246
	oversampling	0.8006	0.5965	0.6836	0.8007	0.5881	0.6754
	downsampling	0.8393	0.6253	0.7166	0.8432	0.6195	0.7116
Maximum Entropy	full-training	0.8899	0.6630	0.7598	0.8887	0.6527	0.7497
	oversampling	0.8869	0.6608	0.7573	0.8770	0.6461	0.7411
	downsampling	0.8452	0.6297	0.7217	0.8363	0.6231	0.7141
Self-training	full-training	0.8958	0.6674	0.7649	0.8938	0.6571	0.7544
	oversampling	0.8929	0.6652	0.7624	0.8877	0.6533	0.7497
	downsampling	0.8631	0.6430	0.7370	0.8542	0.6292	0.7217
CRFs (baseline)		0.8332	0.7172	0.7709	0.8296	0.7152	0.7682
Reserved self-training		0.9194	0.6829	0.7837	0.9134	0.6785	0.7786
Reserved self-training with min. cuts		0.9313	0.6918	0.7939	0.9254	0.6874	0.7888

Figure 2 shows the performance of supervised polarity classifiers for 20 topics based on different imbalanced classification methods. We employed MaxEnt for both self-training and reserved self-training in our subsequent experiments because MaxEnt performed better than SVM. Contrary to the results of Li et al. (2011) in which downsampling approach performed best, in our study, full-training performs at least not bad than downsampling and oversampling. We speculate that these 20 microblog topics involve multiple domains such as political, environmental, and health issues, it would lose some potentially useful information if downsampling method is applied, which induced a bad performance of downsampling. In addition, all the methods perform badly on topic 3, 6 and 11 in which positive sentences account for the major class as shown in Li et al. (2011). There are two possible reasons for these results: (1) training data set of the NLP&CC2012 is imbalanced, the number of negative sentences is 4 times that of positive one, which results in models bias towards the majority class, namely negative sentences; (2) topic 3, 6 and 11 contain much more positive sentences than the others.

Reserved Self-training for Imbalanced Sentiment Classification of Chinese Microblogs

In this subsection, we report the performance of reserved self-training on imbalanced sentiment classification of Chinese microblogs. We reimplemented a model that achieved the best run in the NLP&CC2012 for comparison. It employed Conditional Random Fields(CRFs) to predict unlabeled data, and we treated this model as our evaluation baseline in our experiments.

The entire labeled training corpora is divided into three groups, a labeled training corpus, an unlabeled data set that is actually annotated in order to facilitate the experiments, and a reserved labeled sample set. We perform fivefold cross-validation and use the averaged results as our final estimation. In Figure 3, we can see that reserved self-training performed better than the other methods, especially on topic 3, 6, and 11 in which positive sentences accounted for the major class. A detail comparison of different methods can be found in Table 2. It is worth mentioning that incorporating context information by minimum cuts is able to enhance the performance of our results.

6 Conclusions and Future work

In this study, we focus on the problem of imbalanced sentiment classification of Chinese microblogs. Experiments show that reserved self-training could effectively make use of imbalanced labeled and unlabeled data to achieve better performance with less training data compared with full training, while downsampling and oversampling failed to make improvement. Additionally, combining the context information between different sentences based on minimum cuts is able to revise bad classification. Inspired by active learning,

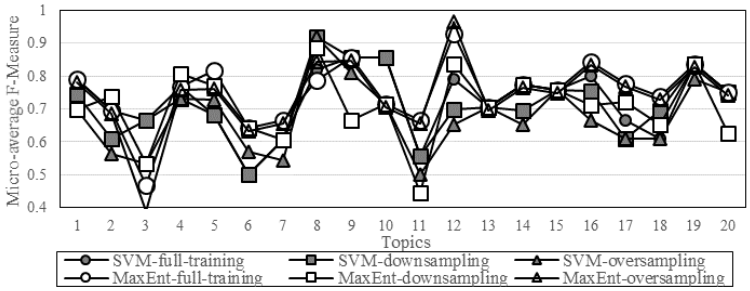

Figure 2: Performances of supervised polarity classifiers for different topics

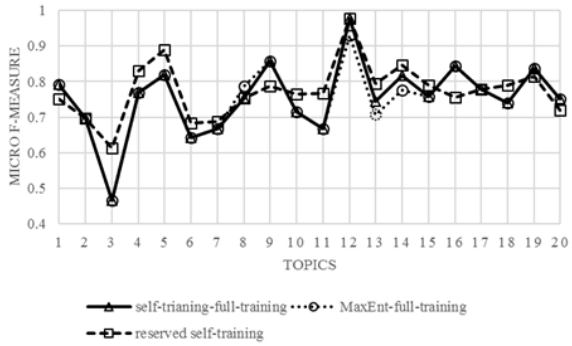

Figure 3: Comparison of different approaches with reserved self-training on imbalanced data

reserved self-training incorporate both the most confident unlabeled samples, together with their predicted labels, and the least confident labeled samples into training set. The classification error can be reduced because the least confident labeled samples would help the model better discriminate different classes. Thus, the selection strategy of reserved self-training can be applied to resolve other problems involving imbalanced binary classification, and not restricted to sentiment classification of microblogs. In the future, we will try to extend this method to address multi-label classification problems.

Acknowledgments

This work was supported by National Natural Science Foundation of China (NSFC) via grant 90924015.

References

Maria-Florina Balcan, Avrim Blum, and Ke Yang. 2004. Co-training and expansion: Towards bridging theory and practice. In *Advances in neural information processing systems*, pages 89–96.

Soumya Banerjee and Nitin Agarwal. 2012. Analyzing collective behavior from blogs using swarm intelligence. *Knowledge and information systems*, 33(3):523–547.

Ricardo Barandela, José Salvador Sánchez, Vicente Garcıa, and Edgar Rangel. 2003. Strategies for learning in class imbalance problems. *Pattern Recognition*, 36(3):849–851.

Gustavo EAPA Batista, Ronaldo C Prati, and Maria Carolina Monard. 2004. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, 6(1):20–29.

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100. ACM.

Ming-Wei Chang, Lev Ratinov, and Dan Roth. 2007. Guiding semi-supervision with constraint-driven learning. *Urbana*, 51:61801.

Olivier Chapelle, Bernhard Schölkopf, Alexander Zien, et al. 2006. *Semi-supervised learning*, volume 2. MIT press Cambridge.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. *J. Artif. Int. Res.*, 16(1):321–357, June.

Nitesh V Chawla, Nathalie Japkowicz, and Aleksander Kotcz. 2004. Editorial: special issue on learning from imbalanced data sets. *ACM SIGKDD Explorations Newsletter*, 6(1):1–6.

Chris Drummond, Robert C Holte, et al. 2003. C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. In *Workshop on Learning from Imbalanced Datasets II*, volume 11. Citeseer.

Yen-Hsien Lee, Paul Jen-Hwa Hu, Tsang-Hsiang Cheng, and Ya-Fang Hsieh. 2012. A cost-sensitive technique for positive-example learning supporting content-based product recommendations in b-to-c e-commerce. *Decision Support Systems*, 53(1):245–256.

Shoushan Li, Zhongqing Wang, Guodong Zhou, and Sophia Yat Mei Lee. 2011. Semi-supervised learning for imbalanced sentiment classification. In *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence-Volume Volume Three*, pages 1826–1831. AAAI Press.

Charles X Ling, Qiang Yang, Jianning Wang, and Shichao Zhang. 2004. Decision trees with minimal costs. In *Proceedings of the twenty-first international conference on Machine learning*, page 69. ACM.

Bing Liu, Minqing Hu, and Junsheng Cheng. 2005. Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of the 14th international conference on World Wide Web*, pages 342–351. ACM.

Jonathan Ortigosa-Hernández, Juan Diego Rodríguez, Leandro Alzate, Manuel Lucania, Iñaki Inza, and Jose A Lozano. 2012. Approaching sentiment analysis by using semi-supervised learning of multi-dimensional classifiers. *Neurocomputing*, 92:98–115.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, page 271. Association for Computational Linguistics.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics.

Ellen Riloff, Janyce Wiebe, and Theresa Wilson. 2003. Learning subjective nouns using extraction pattern bootstrapping. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 25–32. Association for Computational Linguistics.

Russell J Ryan. 2011. *Groundtruth budgeting: a novel approach to semi-supervised relation extraction in medical language*. Ph.D. thesis, Massachusetts Institute of Technology.

Aarti Singh, Robert Nowak, and Xiaojin Zhu. 2008. Unlabeled data: Now it helps, now it doesn't. In *Advances in Neural Information Processing Systems*, pages 1513–1520.

Peter D Turney. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 417–424. Association for Computational Linguistics.

Gary M Weiss. 2004. Mining with rarity: a unifying framework. *ACM SIGKDD Explorations Newsletter*, 6(1):7–19.

Wei Wu and Mari Ostendorf. 2013. Graph-based query strategies for active learning.

Show-Jane Yen and Yue-Shi Lee. 2009. Cluster-based under-sampling approaches for imbalanced data distributions. *Expert Systems with Applications*, 36(3):5718–5727.

Bianca Zadrozny, John Langford, and Naoki Abe. 2003. Cost-sensitive learning by cost-proportionate example weighting. In *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, pages 435–442. IEEE.