一般社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report

Semi-supervised Sentiment Classification in Resource-Scarce Language: A Comparative Study

Yong ren[†] Nobuhiro Kaji[‡] Naoki Yoshinaga[‡] Masashi Toyoda[‡] Masaru Kitsuregawa[‡]

† Graduate School of Information Science and Technology, The University of Tokyo 7-3-1 Hongo,Bunkyo-ku, Tokyo, 113-0033 Japan

‡ Institute of Industrial Science, The University of Tokyo 4-6-1 Komaba, Meguro-ku, Tokyo, 153-8505 Japan

E-mail: † renyong@tkl.iis.u-tokyo.ac.jp, ‡ { kaji, ynaga, toyoda, kitsure}@ tkl.iis.u-tokyo.ac.jp

Abstract With the advent of consumer generated media (e.g., Amazon reviews, Twitter, etc.), sentiment classification becomes a heated topic. Conventional approaches heavily rely on a large amount of linguistic resources, which are difficult to obtain in resource-scarce languages. To overcome this problem, semi-supervised learning (SSL) algorithms have been exploited. However, for the development and variety involved in SSL literature, when people try to adopt SSL approach in practice, they usually confront difficulty in deciding the proper method from many potential candidates. In this study, we conduct empirical evaluation on several representative SSL algorithms in a document-level sentiment classification task for resource-scarce languages (Chinese in our case), and the comparative experiment is carried out using three real datasets. We will describe corresponding theorems, show characteristics and related existing issues for each evaluated algorithm. We believe the other people who interested in exploiting SSL methods could benefit from our experience.

Keyword sentiment classification, semi-supervised learning, comparative study

1. Introduction

Over the last decade, document-level sentiment classification has attracted much attention from NLP researchers; its potential applications include opinion summarization and opinion mining [12]. Most of the existing methods locate sentiment classification as a supervised classification problem and train a reliable classifier from a large amount of labeled data [4, 9, 10, 13]. The main disadvantage of such supervised approaches is that it is quite expensive in both time and labor to annotate a large amount of training data.

Unfortunately, in some languages such as Chinese and Hindi, a sufficient amount of training data is not always available. It is known that the data labeling procedure is quite time consuming, and substantial human labor even linguistics experts are required during this process. Sentiment classification becomes a quite challenging problem for such resource-scarce languages. In order to tackle this problem several studies have adopted semi-supervised learning (SSL). Compared with supervised learning counterparts, the most compelling aspect of SSL approaches is the capability to carry out learning on both labeled and un-labeled data. Unlike labeled data, un-labeled data are much easier to attain, so in SSL methods the dependency on labeled data is highly relieved.

While various SSL algorithms [20] have been proposed recently, there are rare researches on comparative study of SSL approaches. One exception is [8]}, the author measure the performance of four different SSL algorithms in a conventional text topic categorizing task. To the best of our knowledge, there is no previous work that carries out empirical study in document-level sentiment classification where only a few amount of training data are available. In order to get a comprehensive learning on SSL methods, such kind of evaluation is necessary.

In this paper, we focus on exploring the use of SSL measures in building a document-level sentiment classifier under a minimally-supervised setting, where we have only a small number of labeled reviews other than the target reviews that we want to classify. Several typical methods including label propagation (LP) [21], modified adsorption (MAD) [16], transductive Supported Machines (TSVM) [5] and spectral graph transducer (SGT) [7] are evaluated. We will explain the ideas behind them, show the relationship among them, and present their pros and cons in our task.

The main contribution of our work is: we evaluate the use of several representative SSL algorithms on document-level sentiment classification in a resource-scarce language. The comparative analysis are conducted are via experiment on real Chinese reviews

This article is a technical report without peer review, and its polished and/or extended version may be published elsewhere.
Copyright ©20●● by IEICE

from three different domains. And we found after parameter-tuning, SSL algorithms(TSVM and SGT) who own excellent supervised learning base can outperform those random-walk based counterparts(LP and MAD).

Though four SSL methods are evaluated in our study, for the similarity exists in SSL literature. We believe our work is helpful for those people also who try to make use of SSL approaches in their research, especially, they will get a clearer understanding on the issues in exploiting SSL approaches in practice.

The rest of this paper is organized as follows: Section 2 introduces related works; Section 3 explains the theory and applying usages of graph-based SSL methods explored in our study in depth. Section 4 evaluates those algorithms. Section 5 concludes this study and discusses future direction.

2. Related Work

For sentiment classification task, several researchers attempted to solve this task in languages without abundant training instances [3, 15, and 19]. In what follows, we briefly introduce those studies.

[3] used transductive SVMs [5] to exploit unlabeled reviews in a document-level sentiment classification task. Basically their work is divided into three steps: firstly they perform spectral clustering to identify unambiguous reviews from unlabeled reviews. Second, they employ active learning to label the remaining ambiguous reviews. Finally, they use the resulting labeled reviews and the remaining unlabeled reviews to train a transductive SVM classifier. This study assumes manual intervention in the active learning step. TSVMs is also one baseline system in our study.

[19] adopted a graph-based propagation approach called Potts model [18] to solve a sentence-level sentiment classification task. Similar to label propagation we explored in our study, Potts model uses the relationship among instances, and each instance arrives a probability state through the process of propagation until the whole graph stabilizes. We should mention that the motivation of their study is not to obtain high classification performance in a minimally-supervised setting but to make use of intra- and inter-document evidences in sentence-level sentiment classification. The usefulness of a graph-based semi-supervised algorithm in a minimally-supervised setting remains to be investigated.

[15] exploited Modified Adsorption (MAD) [16] for "tweets" polarity categorization. The MAD algorithm is an improved version of label propagation. And we will show how the improvement can play positive role in our work. For the social networking characteristics of Twitter, the authors build the similarity graph by considering more factors available, e.g., following/follower relationship. They took several labeled data (seeds) for MAD; the labels of these seeds come from special corpus or delicate training methods. Different from this work, the seeds used in our work are chosen randomly from the data need to be classified. And we measure those SSL methods without using any kind of outside resources. So we believe our work are more general and reflect real performance of related algorithms in reality.

3. Evaluated Methods

3.1. Methods Overview

Broadly speaking, semi-supervised learning include two subclasses: semi-supervised classification and constrained clustering [20]. Here we use the collocation "semi-supervised learning"(SSL) to refer to semi-supervised classification.

SSL can learn from both labeled data and un-labeled data. Typically we assume there are much more unlabeled data than labeled data. The capability of SSL is learning the label of those un-labeled data from labeled data. One common solution is exploring the latent structure in the whole data, other than merely the labeled data just like supervised leaning does.

SSL can be divided into several groups based on the data structure they exploit. Two common groups are graph-based SSL and transductive support vector machines (TSVM). In the following part in this section, we will describe the typical algorithms especially hallmark the in detail and explain the relationship among them. In section \ref{experiment}, we will show the potential effect of characteristics in each algorithm.

3.2. Graph Based Semi-Supervised Learning

Given the similarity graph $G=\{V, E, W\}$ consists of vertices V, edges E, and an $n* n$ weighted similarity matrix $W = [w_{ij}]$, where $n=|V|$. The edge weight w_{ij} is calculated by similarity score between review i and review j. One common supposition is graph-based SSL is the larger similarity score w_{ij} between x_i and x_j is, the more likely they own the same label.

The first graph-based SSL we explore is Label propagation [21]; it owns a lot of advantages including convergence and a well defined objective function. Move over, it has been successfully employed in several NLP tasks, such as sentiment lexicon induction [14] and word sense disambiguation [11].

Mathematically speaking, LP aims to minimize the following objective function [14], where *l* is the number of labeled data, *u* is the number of unlabeled data, and y_i is a binary label that takes 1 if the vertex is positive and 0 if the vertex is negative.

$$\sum_{i,j=1}^{l+u} w_{ij}(f(x_i) - f(x_j))^2 \text{ subject to } f(x_i) = y_i \text{ for } i = 1, 2, \ldots, l$$

The solution to this problem satisfies the following stationary conditions:

$$f(x_i) = y_i, i = 1 \ldots l \quad f(x_j) = \frac{\sum_{k=1}^{l+u} w_{jk} f(x_k)}{\sum_{k=1}^{l+u} w_{jk}}$$

Intuitively, label propagation seeks *f(i)(i=l+1,l+u)* that satisfies in this Equation in iterative manner.

Adsorption [1] is modification of LP, both the theory and executing procedure are resembled to LP. It is also a graph-based SSL method, each vertex in the graph still takes two roles: one is to learn its own label form its neighbors, the other one is to propagate its label information to its neighbors. The key differences are: firstly, the labeled vertices in the similarity graph are not allowed to adjust to the original value as things happen in LP. One main motivation of this strategy is to empower this approach with the capabilities with noise or un-reliable labels. And "Shallow" vertices are added as adsorbing ends. Furthermore, adsorption brings inject, continue and give up probabilities into the label propagating process for each vertex. In other words, adsorption could be taken as controlled label propagation. It could be adapted to diverse graphs in more flexible way, at the price of model complexity. We will get a clear understanding on this point when we compare adsorption and LP in the following experiment section.

Modified adsorption (MAD) [16] is a further enhanced version of adsorption. The main motive of MAD is to alter original adsorption algorithm so that it could own an objective function, then the algorithm could gain optimal output through optimization methodologies. Moreover, in MAD the selection of inject, continue and give up probabilities are further explored. Detailed computing formula could be found in the work [16]. The objective function is constructed by take three constrains into account: the result of labeled instances should be kept consistent with the corresponding inherent value as much as possible. Secondly, the higher the weight between two vertices the closer the label value they share. At last, the whole result should be as uninformative as possible. All the three considerations are combined together, and three dependent parameters are used to measure the importance of those three factors. Moreover, the three parameters are just co-related with the three probabilities proposed in the Adsorption method respectively. For simplicity, when we make use of MAD in our task, we take the related parameters as the weight to emphasize the importance of different controlled factors. We will explain it in detail in the experiment section later.

Spectral Graph Transducer [7] (SGT) is proposed by T.Joachims in 2003. Different from LP and MAD, SGT is based on graph partitioning. It can be viewed as transductive version of k nearest-neighbor (kNN) classifier [7]. This process can be expanded formally as follows:

$$\sum_{i}^{n} \sum_{j \in kNN(x_i)} y_i y_j \frac{w_{ij}}{\sum_{m \in kNN(x_i)} w_{im}}$$

yi and y_j are binary label value for annotated data. As a matter of fact, from the formula (3), we can clearly see that SGT also tries to keep the fundamental supposition of graph-based SSL campaign, near neighbors share the same label. If we treat the labeled data as sources and ends, the optimization problem depicted in formula (3) can be resolved by adopting s-t graph cut paradigm [7].

However, besides the two conventional postulates: low training error and corresponding high performance inductive learner that many semi-supervised learner own, SGT takes a further step to make pos/neg ratio have the same expected value in the training and in the test set. Then the author outlines SGT as normalized graph cut with constrains, and piggyback spectral graph theory to find the optimal result. The purpose is to cure the degenerate cuts existing in s-t graph cut algorithm, in other words, SGT takes measures to avoid the appearance of un-balanced classification.

3.3. Semi-supervised Support Vector Machines

Semi-supervised Support Vector Machines (S3SVMs) was originally called Transductive Support Vector Machines (TSVMs), it is firstly proposed in [5] and then refined in [6] and [2]. As the name suggested, TSVMs is a transductive brother of SVMs. We can see clearly from the graph, the purpose of SVMs is build a optimal hyper-plane to divide the data space with the help of labeled data points; on the other hand, TSVMs can find a classifier by considering both labeled and un-labeled data points. And the difference between S3VMs and T3VMs is located on the final purpose is to find a general hyper-plane for the whole data space or not

Mathematically, the objective function in S3SVMs is listed in formula:

$$\min_{w,b} \sum_{i=1}^{l} \max(1-y_i(w^T x_i+b),0)+\lambda_1 \|w\|+\lambda_2 \sum_{j=l+1}^{l+u} \max(1-\left|w^T x_j + b\right|,0)$$

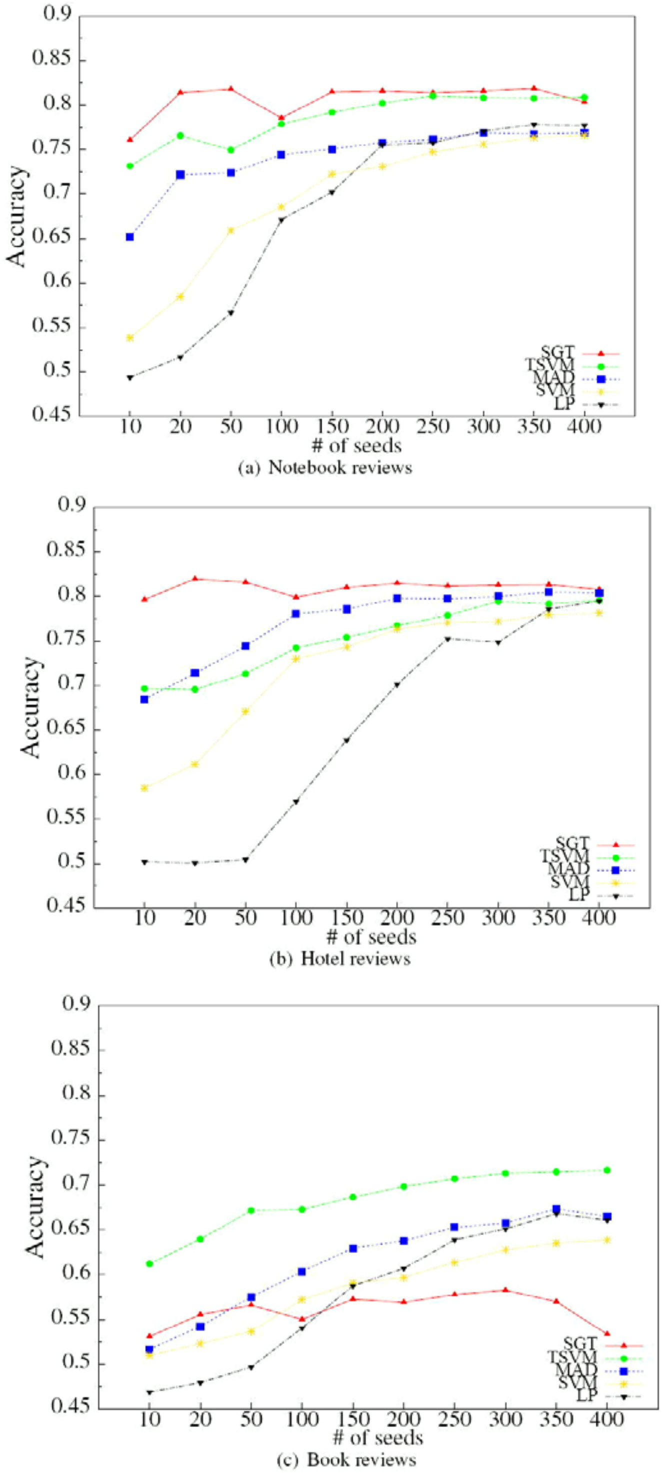
In fact, the part former part before $\lambda_2$ is the principle of SVMs; simply speaking, it tries to classify the training instances, at the same time, maximize the distance between classifier and the training data. The latter part after $\lambda_2$ distinguishes it from SVMs, which focus on building a hype-rplane to avoid data region with high dense.

Finally, from the description above, though the principles involved are different, there is something consistent in various SSL methods. All of them suppose data distribution (include both labeled and unlabeled) can be helpful to identify the categorization of data, specifically, the densely clustered data points opt to below to the same class. Based on this suppose, diverse ideas are developed: propagate label through the similarity graph makes the data densely connected share the same label; graph cut is dividing the whole graph into densely connected subgraphs; S3VMs tries to find a hyper-plane to go through sparse region.

4. Evaluation and Discussion

4.1. Data Set

The dataset we used is ChnSentiCorp (de-duplicate version)[1]. It consists of reviews from three different domains: notebook, hotel, and book. The sentiment polarity of each text has been manually labeled. In each domain, we randomly selected 400 reviews as test dataset. The ratio between positive and negative reviews in the test data is 1:1, so that random baseline achieves accuracy of 0.5. Then we randomly selected 400 different reviews as validation data for tuning the parameters in each algorithm. Finally, we divided the remaining reviews into two parts: labeled seeds unlabeled data. The number of labeled seeds is varied from 10 to 400 to control the amount of supervision.



(a) Notebook reviews

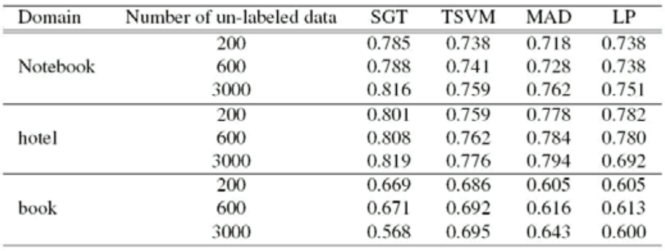
(b) Hotel reviews

(c) Book reviews

Figure 1

[1] http://www.searchforum.org.cn/tansongbo/corpus-senti.htm

Table 1 Influence of un-labeled data



Domain	Number of un-labeled data	SGT	TSVM	MAD	LP
Notebook	200	0.785	0.738	0.718	0.738
	600	0.788	0.741	0.728	0.738
	3000	0.816	0.759	0.762	0.751
hotel	200	0.801	0.759	0.778	0.782
	600	0.808	0.762	0.784	0.780
	3000	0.819	0.776	0.794	0.692
book	200	0.669	0.686	0.605	0.605
	600	0.671	0.692	0.616	0.613
	3000	0.568	0.695	0.643	0.600

In order to decrease the effect of seed selection, we conduct a twenty-round experiment for the same number of labeled seeds. In each round, different labeled seeds are used. We report the average classification accuracy as the final result.

In order to get a more comprehensive interpretation, we also add SVMs which is a widely utilized supervised classification algorithms as a baseline.

We use SVMlight[2] as implementations of SVMs and TSVMs used in our experiments; we employ Junto[3]as implementation of LP and MAD; and we adopt SGTlight[4] as implementation of SGT.

4.2. Experiment and Result

The comparison of classification performance is shown in Figure 1; the vertical axis indicates the accuracy, while the horizontal axis indicates the number of labeled seeds. All the algorithms (except the parameter-free LP) are best tuned. The performance is based on the test dataset.

At first, the performance of LP is bad when the size of annotated instances is small. Possible culprit locates in the imperfect structure of similarity graph. Some common phrases (sentiment features) extracted take the role as undesired bridge to connect positive instances and negative instances. More badly, when reliable label sources are scarce, LP will make wrong decision. We should note that "mis-link" phenomenon among instances is a common problem. How to take effective measure to combat the issue is still undergoing.

As the improved version of LP, MAD can over-perform LP in most of cases, especially when the size of available labeled seeds in limited. The possible reason is MAD could tackle imperfect similarity graph with flexibility. After parameter tuning, the role of labeled data is emphasized properly, at the same time, the propagation behavior get appropriate control. The hint given here is when we cannot construct desirable graph, take the strategy to control propagating label across the graph could be helpful.

Moreover we can see that, with the increase of labeled instances, the performances of all of methods are improved. For MAD and LP when more seeds take part in the label propagating process, unlabeled vertices could get more reliable sources so that MAD or LP could become more confident to decide the label one specific vertex belongs to. One the other hand, for SVM and TSVM, the increase of labeled instances means more labeled data are available, so the classifiers perform better.

Finally, all of those approaches do not perform well in book domain. Because in book reviews people would discuss various aspects including the story, the writing style of author, the figures in the book and even the publisher. Sometimes the sentiment those aspects are not consistent. [17] reported similar findings on movie reviews.

Furthermore, we also explore the influence of unlabeled data. Specifically speaking, we will present what will happen if we change the number of un-labeled data. We believe this is an important aspect when people adopt semi-supervised learning algorithms in practice. For instance, if we cannot find the emergence of better result as we add more un-label data or even side-effect appears, we can take the measure to limit the number of un-labeled instances. Moreover, our investigation could also be helpful for understanding of SSL approaches.

In order to check the role of un-labeled data, we fix the number of labeled seeds to 200 (these labeled data are the same as the experiment before). Totally three different un-labeled data size is used: 200, 600 and 3000 respectively. We also take the measure of 20 rounds experiment (un-labeled data varies with group of labeled data) and report the average result. For SVM is a supervised leaning, it is excluded in this test.

The experiment result is listed in Table 1.

We could find that in most cases, the increasing of un-labeled data does not promote the performance in LP. The reason is with the more un-labeled data added, more links are built but some of them are mis-links. Those mis-links could have side-effect on the final performance. For example, in hotel review, the result has obvious decrease. Benefit from the flexibility, MAD can digest the noise brought by un-labeled data. However, the performance is not improved much by contrast to the increase of un-labeled instances. Similarly, we could also get a conclusion that the promotion is not significant in TSVM. More un-labeled data could help TSVM to find a

[2] http://svmlight.joachims.org/
[3] https://github.com/parthatalukdar/junto
[4] http://sgt.joachims.org/

confident classifier, but it also means more noise. From our investigation, when we exploit SSL algorithms, we should not anticipate too much on enhancing classifying performance by investing more un-labeled instances.

In this paper we evaluate the use of several representative semi-supervised learning algorithms: lable propagation, modified adsorption, transductive support vector machines and spectral graph transducer for a document-level sentiment classification task in resource-scarce languages. We describe the theorem involved in them and explained the relationship among them. Especially, we analyze the existing issue in LP and how the improvement done in MAD could tackle the issue. We believe our experience is important in exploiting SSL methods.

We present the parameter tuning process exhaustively and investigate the impact on performance. All the experiments are done on real Chinese reviews classification in three domains.

In the future, we will investigate SSL algorithms in sentiment classification on a large scale dataset. We will pay much attention of algorithm complexity and efficiency. We will also explore the usages of those algorithms in free text such as tweets and so forth.

Reference

[1] D. Ravichandran, and M. Aly. Video suggestion and discovery for youtube: taking random walks through the view graph. In Proceedings of the 17th international conference on World Wide Web, WWW' 08, pages 895–904. ACM, 2008.

[2] K. P. Bennett. Combining support vector and mathematical programming methods for classification. In B. Sch¨olkopf, C. J. C. Burges, and A. J. Smola, editors, Advances in Kernel Methods: Support Vector Learning. MIT Press, Cambridge, MA, 1999.

[3] S. Dasgupta and V. Ng. Mine the easy, classify the hard: a semi-supervised approach to automatic sentiment classification. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2, ACL '09, pages 701–709, 2009.

[4] M. Gamon. Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. In COLING, pages 841–847, 2005.

[5] T. Joachims. Transductive inference for text classification using support vector machines. pages 200–209. Morgan Kaufmann, 1999.

[6] T. Joachims. Transductive inference for text classification using support vector machines. In ICML, pages 200–209, 1999.

[7] T. Joachims. Transductive learning via spectral graph partitioning. In ICML, pages 290–297, 2003.

[8] C. Lanquillon. Learning from labeled and unlabeled documents: A comparative study on semi-supervised text classification. In Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery, PKDD '00, pages 490–497, London, UK, UK, 2000.

[9] S. Matsumoto, H. Takamura, and M. Okumura. Sentiment classification using word sub-sequences and dependency sub-trees. In PAKDD'05, pages 301–311, 2005.

[10] T. Mullen and N. Collier. Sentiment analysis using support vector machines with diverse information sources. In In Proceedings of Conference on Empirical Methods in Natural Language Processing, 2004.

[11] Z.-Y. Niu, D.-H. Ji, and C. L. Tan. Word sense disambiguation using label propagation based semi-supervised learning. In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, ACL '05, pages 395–402, 2005.

[12] B. Pang and L. Lee. Opinion mining and sentiment analysis. Found.Trends Inf. Retr., 2:1–135, January 2008.

[13] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In EMNLP '02: Proceedings of the ACL-02 conference on Empirical methods in natural language processing, pages 79–86. Association for Computational Linguistics, 2002.

[14] D. Rao and D. Ravichandran. Semi-supervised polarity lexicon induction. In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, EACL '09, pages 675–682, 2009.

[15] M. Speriosu, N. Sudan, S. Upadhyay, and J. Baldridge. Twitter polarity classification with label propagation over lexical links and the follower graph. In Proceedings of the First Workshop on Unsupervised Learning in NLP, EMNLP '11, pages 53–63,Stroudsburg, PA, USA, 2011.

[16] P. P. Talukdar and K. Crammer. New regularized algorithms for transductive learning. In Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases: Part II, ECML PKDD '09, pages 442–457, Berlin, Heidelberg, 2009.

[17] P. D. Turney. 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, ACL '02, pages 417–424, 2002.

[18]F. Y.Wu. The potts model. Rev. Mod. Phys., 54(1):235–268, Jan 1982.

[19] B. Q. Yanyan Zhao and T. Liu. Integrating intra- and inter-document evidences for improving sentence sentiment classification. ACTA AU-TOMATICA SINICA, 36(10):1417–1425, 2010.

[20] X. Zhu. Semi-supervised learning literature survey, 2006.

[21] X. Zhu and Z. Ghahramani. Learning from labeled and unlabeled data with label propagation. Technical report, 2002.S