Measuring Feature Distributions in Sentiment Classification

Diego Uribe

Instituto Tecnológico de la Laguna
División de Posgrado e Investigación
Blvd. Revolución y Calz. Cuauhtémoc, Torreón, Coah., MX
diego@itlalaguna.edu.mx

Abstract. We address in this paper the adaptation problem in sentiment classification. As we know, available labeled data required by sentiment classifiers does not always exist. Given a set of labeled data from different domains and a collection of unlabeled data of the target domain, it would be interesting to determine which subset of those domains has a feature distribution similar to the target domain. In this way, in the absence of labeled data for a particular target domain, it would be plausible to make use of the labeled data corresponding to the most similar domains.

1 Introduction

A huge volume of opinionated text is nowadays a common scenario in the Web. It is precisely this enormous information that represents a great sandbox for developing practical applications for the industry and the government. For example, before buying a product or service, people look for someone else's experiences about the product on the web, i.e. opinions. The same occurs with services provided by government or industry. In this way, persons and organizations are interested in the use of opinionated text for decision optimization: persons taking care of their money, and organizations getting significant feedbacks for improving their products or services [2].

This increasing interest in the use of opinionated text demands automated opinion discovery and classification systems. Sentiment classification is basically a text classification task in which, instead of assigning one topic to the text, the attitude of the reviewer with respect to the text is determined. As emotions are not only described in many ways but also differently interpreted by persons, sentiment classification is one of the most defiant problems in computational linguistics.

To cope with the linguistic challenge demanded by the automatic classification of emotions, diverse machine learning techniques have been applied. Most of the current work has been concerned with the use of supervised classification techniques which require large amounts of labeled training data [12],[5]. However, as labeling data is time-consuming, in some domains only limited training data exists. As training data might be limited, it could be fruitful revealing to

I. Batyrshin and M. González Mendoza (Eds.): MICAI 2012, Part II, LNAI 7630, pp. 69–80, 2013.

investigate how existing resources could be used to train a classifier in different domains.

Since sentiment in different domains can be expressed in very different ways, training a classifier using data from one domain may fail when testing against data from another, different domain. In other words, we face greater difficulty when the available training instances differ from the target domain. Aue and Gamon illustrate how the accuracy of a classifier trained on a different domain drops significantly compared to the performance of a classifier trained on its own native domain [1]. Thus, to determine which subset of outside domains has a feature distribution most similar to the target domain is of paramount importance.

In this paper, it is proposed a framework for estimating which subset of labeled corpora domains has a feature distribution similar to the unlabeled target domain. Our approach gets its inspiration from the task of identifying the author of a given document -authorship attribution [6]. To be more specific, given a set of labeled data from different domains, a domain's profile is created by information extracted from the reviews corresponding to each domain. Once the profiles have been defined, an unlabeled review, corresponding to the target domain, is associated with the domain whose profile is most similar to the review. Thus, the specification of the linguistic terms to be extracted from each review is crucial for the definition of each domain's profile as well as the similarity metric used to identify unseen reviews.

We evaluate the proposed method with a data collection of several domains [17]. In particular, we used data from 4 different domains: Books, Movies, Music, and Phones. The feature distribution for each of the held-out domains is generated using profiles created on data from the other three domains. We then verify the feature distribution with classifiers based on the profile of each of the three domains and tested on the held-out domain (target domain). The results of the experimentation show how a feature distribution based on the semantic orientation of available datasets is a plausible option for the classification of unlabeled reviews.

The rest of the paper is organized as follows. The next section makes a brief review of of related work on the problem of domain adaptation in sentiment analysis. Section 3 describes in detail the proposed method for the creation of each domains profile and the metrics for computing the similarity between unlabeled reviews and domains profiles. Then, the setup and results of the experimentation are exhibited and discussed in section 4. Finally, conclusions are given in section 5.

2 Related Work

There is substantial work in sentiment classification using different strategies. Indeed, diverse machine learning techniques have been applied to automatically classify opinions. In this paper we briefly describe some of the preceding works into customizing sentiment classifiers to new domains we briefly mention some interesting works. One of the first works in this specific topic was carried out by

Aue and Gamon [1]. Their work is based on multiple ways of fusion of labeled data from other domains and unlabeled data from the target domain. The best results were obtained with an approach based on bootstrapping techniques. It is important to notice our investigation was initially proposed in this previous work.

Yang et al. [20] also deal with domain adaptation in their opinion detection system. Subjective and objective sentences from movies and products reviews were used as the training data for the task of opinion detection across the blog posts. Multiple feature vectors based on n-grams were used in the experimentation. Only those features highly ranked in both domains were considered as generic features for the classification task. Likewise, Blitzer et al. [4] cope with the domain adaptation problem by extending an algorithm for sentiment classifier by making use of *pivot features* that relate the source and target domains. This relationship is defined in terms of frequency and mutual information estimation.

Shou-Shan et al. [14] propose an interesting algorithm for multi-domain sentiment classification based on the combination of learners for specific domains called *member classifiers*. The combination of these member classifiers is done according to two types of rules: fixed and trained rules. The purpose of the combination process is to obtain and to make available global information to the final classifier.

3 Proposed Method

In this section, we describe in detail our approach to estimate the subset of different domains with a feature distribution similar to the target domain. As we previously said, our approach gets inspiration from the task of identifying the author of a given document, that is, authorship attribution [16]. As Figure 1 shows, there are two key elements in the framework. First, the creation of each domains profile requires the definition of the linguistic terms to be extracted from each review as well as the associated information of each term, i.e. frequency, semantic orientation. Second, the similarity metric used to get a clearer view on how distant the unlabeled reviews are from each previously defined profile. We analyze in this work two different ways to define a domain profile: a domain profile based on the presence of the subjective terms among diverse domains, and the semantic orientation of the subjective terms used as an alternative way to portray the profile of a particular domain.

3.1 Domain Profile

The first step to estimate the similarity distribution among domains is the creation of each domain's profile. Thus, to determine the granularity of the linguistic terms to be extracted from each review is crucial for the definition of a profile. Most of the previous work in sentiment classification makes use of n-grams models for the representation of the reviews ([12], [13], [20]). Also, another common

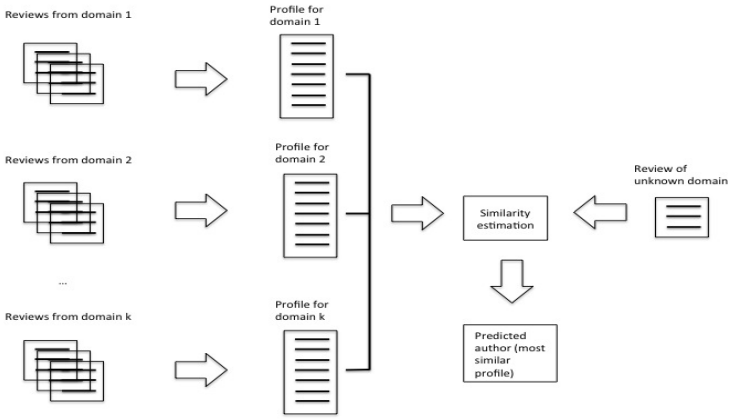

Fig. 1. Domain recognition

scenario is the use of linguistic patterns for the extraction of specific linguistic terms [19]. Since these patterns are based on adjectives and adverbs as good indicators of subjectivity and opinions, they represent a plausible alternative to look for an optimal subset of features. We use in this work a combined approach: a phrase pattern-based method which relies on bigrams and a small set of particular rules proposed by [15].

In fact, our method extracts bigrams as a way to incorporate elemental context in the analysis. But we don't extract any bigrams. We extract only those bigrams matched by the set of rules proposed by [15]. The key point in these rules is the consideration of the context that surrounds the subjective words [12]. For example, phrases such as *"nice"* and *"not very nice"* denote opposite emotions. Thus, the rules are basically applied to adjectives and adverbs and operate against only the three leading contextual word positions. It is also important to observe that the application of the rules is carried out in sequential order. Table 1 shows the rules along with some examples.

In this way, in order to obtain the subjective expressions matched by the linguistic patterns, the set of sentences corresponding to each review was submitted to a tagger: The Stanford Tagger [18]. Thus, a particular *domain profile* is the set of subjective bigrams extracted from their corresponding reviews.

To give a more formal definition, let d represents a specific domain. Then, a *domain profile (dp)* is defined as follows:

$$dp = \{t \mid t \text{ is a subjective term extracted from each } r \text{ where } r \in d\} \quad (1)$$

3.2 Similarity Distribution

To estimate the similarity domain distribution a new unlabeled review and a domain profile are compared. A straightforward and common way to estimate the

Table 1. Steck's patterns

Rule	Example
Three leading positions preceding an adjective and attempts to match negations	*is not very* ***nice***
To match the first leading word that is not a preposition, conjunction, or punctuation mark	*movie was just* ***bizarre***
Applies when the trailing position is identified as a noun	***good*** *actor*

similarity between documents represented by vectors is based on the number of overlapping terms where the Dice, Cosine, and Jaccard coefficients are common metrics. Also, a recent interesting work on social bookmarking makes use of the coefficients to measure the similarity between users of portals for sharing scientific references amongst researchers [8]. However, we observe in this work from a different perspective the intersection of the linguistic terms between different domains and new unlabeled reviews. Two schemes are particularly analyzed: one is based on authorship attribution and the other one relies on semantic orientation.

Presence. In the base model of authorship attribution the use of normalized and un-normalized number of overlapping terms is commonly used as similarity measures [10]. However, we focus our attention on an interesting alternative to the base model for authorship identification suggested by Escalante [6]. Besides considering the number of overlapping terms, the basic idea is to assign a particular weight to each term according to their use across the profiles for different domains. This is precisely the approach that has been adapted in our work to estimate the similarity domain distribution.

Now, if we have k different domains, a domain profile, represented as dp_k, is defined for each particular domain. Then, the weight associated to each t common term between an unseen review and a particular domain profile is defined as follows:

$$w_t = \frac{1}{\sum_{i=1}^{k} at(t, dp_i)} \tag{2}$$

where $at(t, p_i)$ represents a *presence* function defined as:

$$at(t, dp_i) = \begin{cases} 1 & \text{if } t \in dp_i \\ \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

These expressions show how prolific terms among different domains are penalized with an irrelevant weight, whereas those terms whose occurrence is not so common among the profiles strongly influences the assignation of an unlabeled

review to a particular domain. Thus, the similarity measure between a domain profile and an unseen review is defined as follows:

$$sim(r, dp_k) = \frac{1}{n} \times \sum_{i=1}^{n} w_i \tag{4}$$

where n represents the number of terms that overlap between the domain profile dp_k and the unseen r review. In other words:

$$n = r \cap dp_k \tag{5}$$

As we can see, the similarity value in (4) is not only obtained with the total of the common terms' weights. This total is normalized by n overlapping terms in order to reduce the impact of a review with a large number of subjective terms. Thus, the unseen r review will be assigned to the domain profile with the largest similarity value among k different domains.

Semantic Orientation. An alternative approach to estimate the similarity domain distribution is the incorporation of the semantic orientation of the linguistic elements used in the representation of the reviews. The classic Turney's work [19] makes use of a search engine to determine the semantic orientation of the extracted phrases by using PMI between the phrase and polarity words. However, since opinion lexicons are resources that associate words with sentiment orientation, the use of opinion lexicons is also another option for the estimation of the semantic orientation of the terms.

Basically, there are two alternatives to incorporate semantic lexicons in sentiment analysis. First, the use of SentiWordNet (SWN), an extension of WordNet with information about the sentiment of words where each synset has a positive, negative and objectivity score [7], is increasingly common in sentiment analysis. For example, Keefe makes use of SWN in order to propose feature selection methods to reduce the huge number of features found in opinion corpora [9]. Manually created opinion lexicons are another alternative. [12] uses this option with a group of students where they chose the polarity of the terms. In this way, the document's polarity is given by counting positive and negative terms.

In this work we don't make use of SWN and don't manually create a semantic lexicon either. In our phrase pattern-based method, we automatically construct two semantic lexicons for each domain to determine the semantic orientation of each subjective term: the positive and the negative semantic lexicons. The positive lexicon (PL) is automatically created with the whole set of extracted terms corresponding to the positives reviews. In a similar way, the negative semantic lexicon (NL) is also defined. Thus, the semantic orientation (so) of each particular subjective term (t) is estimated according to:

Let ρ be the number of times t occurs in PL

and η be the number of times t occurs in NL,

the degree of subjectivity is described as:

$$\mu = \rho - \eta \tag{6}$$

Then, the semantic orientation associated with each common term corresponding to a particular domain profile (dp) is defined as:

$$so(t, dp) = \begin{cases} 1 & \text{if } \mu > 0 \\ -1 & \text{if } \mu < 0 \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

Once the semantic orientation of a subjective term has been stated, we can now to determine the weight associated to each common term across different domain profiles. Assuming we have k different domains and its corresponding domain profile dp_k, the weight associated to each common term between a an unseen review and a particular domain profile is defined as follows:

$$w_t = \sum_{i=1}^{k} so(t, dp_i) \tag{8}$$

This expression shows how the weight of a term is determined by the subjectivity that such term entails across diverse domains rather than a specific domain. Terms with different polarity across domains (i.e. positive sentiment in one domain and negative sentiment in another) exhibit an irrelevant weight, whereas those terms with regular polarity among the profiles strongly influences the assignation of an unlabeled review to a particular domain. Thus, the similarity measure between a domain profile and an unseen review is defined as follows:

$$sim(r, dp_k) = |\sum_{i=1}^{n} w_i| \tag{9}$$

As we can see, the similarity value in (9) is not only obtained with the total of the common terms' weights. The absolute value of this total is used in order to emphasize the magnitude of the subjectivity that a review entails. Thus, the unseen r review will be assigned to the domain profile with the largest similarity value among k different domains.

As we can see, the similarity value in (9) is obtained with the total of the common terms' weights in order to emphasize the magnitude of the subjectivity that a review entails. Thus, the unseen r review will be assigned to the domain profile with the largest similarity value among k different domains.

4 Experimental Evaluation

We use for our experimental trial the collection of Epinions reviews[1] developed by Taboada et al. [17]. From the whole collection, four domains were selected:

[1] `www.epinions.com`

books, movies, music and phones. There are 50 opinions per category, and since there are 25 reviews per polarity within each category, the baseline accuracy for each domain is 50%. The set of sentences corresponding to each review was submitted to the Stanford Tagger [18] for the specification of the subjective terms to be extracted from each review.

4.1 Similarity Distributions Results

In order to observe the subset of domains with a feature distribution similar to the target domain, each of the selected domains play the role of target domain. For example, to obtain the feature distribution similar to Books, the set of opinions corresponding to Books play the role of unseen reviews to be identified across the Movies, Music, and Phones domains.

Two methods have been described in Section 3 to estimate such similarity distributions. The results of the first method based on the presence of the subjective expressions across the different domain's profiles are shown in Figure 2. The results are displayed in such a way that the column corresponding to each target domain is the lowest one. We observe for example how Music reviews exhibits a feature distribution more similar to Movies than Books.

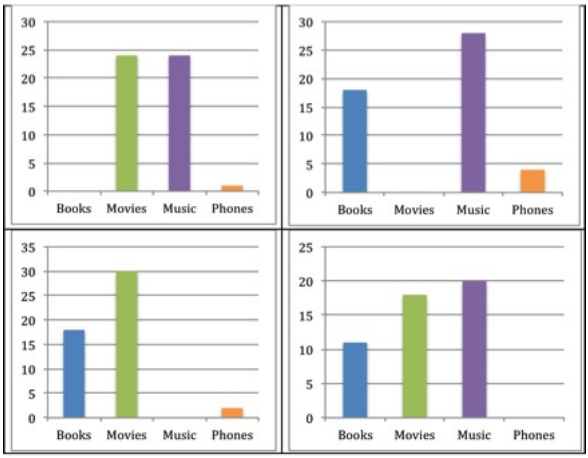

Fig. 2. Presence distribution

In the same way the results of the second method based on the semantic orientation of the subjective expressions across the different domains' profiles are shown in Figure 3. Also, the results are displayed in such a way that the column corresponding to each target domain is the lowest one. In this case we observe for example how Movies reviews exhibits a feature distribution more similar to Books than Music.

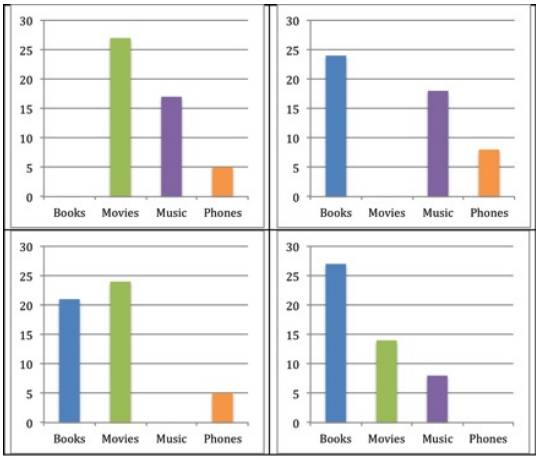

Fig. 3. Semantic orientation distribution

4.2 Analysis and Discussion

By observing the similarity distributions displayed in Figures 1 and 2, we can appreciate some coincidences and differences. For example, we can see how Movies show a feature distribution more similar to Books in both methods, even tough the results in the semantic orientation distribution are more evident. The same occurs when we obtain the distribution for Music: we can see how Movies show a feature distribution more similar to Music in both methods, but the results in the distribution based on presence are more palpable.

On the other hand, the distributions for Movies and Phones differ. For example, we can see how Music show a feature distribution more similar to Movies in the distribution based on presence, but the results in the semantic orientation distribution show Books as the most similar dataset. And the same occurs in Phones. We can see how Music show a feature distribution more similar to Phones in the distribution based on presence, but the results in the semantic orientation distribution show Books as the most similar dataset.

Once we obtained the similarity distributions for each target domain, we want to corroborate such distributions. In other words, we want to observe if, for example, Movies reviews represent a better option to classify Music rather than Books reviews. In order to carry out this corroboration, we evaluate the performance for each target domain using a classifier based on the common features between the unseen reviews of the target domain and each of the foreign domains. We implement two different methods to evaluate the performance.

The first experiment is based on 5-fold cross-validation and the use of support vector machines (SVM). As we know, SVM is a hyperplane classifier that has proved to be a useful approach to cope with natural text affairs [12]. Table 2 shows the SVM classifier accuracy for each target domain. Table 2 exhibits how the use of Movies reviews (60.00%) represent a better option to classify Music

rather than Books reviews (56.00%). Taking into account that the baseline accuracy for each domain is 50%, the performance for unseen reviews from Music and Movies is poor, whereas for unseen reviews from Books and Phones a plausible performance is obtained.

Table 2. Scores obtained with SVM

	Books	**Movies**	**Music**	**Phones**
Books		69.39%	65.31%	61.22%
Movies	42.00%		48.00%	32.00%
Music	56.00%	60.00%		60.00%
Phones	79.60%	79.60%	77.60%	

The second evaluation method portrays more optimist results. This alternative evaluation is based on the use of the two semantic lexicons created for the definition of each domain profile with a different approach: instead of using such domain profiles to identify unseen reviews, we use them to classify the unseen reviews. In this way, we determine the semantic orientation of each subjective term to represent an unseen review as a feature set that denotes features intensity rather than features presence. The classification of the unseen review r is then defined in a similar way to the expression (9).

Table 3 shows the accuracy results of classifiers based on semantic orientation for each target domain. In this Table 3 we see for example how the use of Books reviews represents a better option to classify Movies rather than Music reviews. Likewise, by comparing Table 2 and Table 3, we see how the results obtained with classifiers based on the semantic orientation of the common features also represent a plausible alternative.

However, the most important point has been to observe how the information provided by the similarity distributions (Figure 2 and Figure 3) allows selecting domains with a feature distribution similar to the unlabeled target domain. To cope with the problem of domain adaptation is important to make a deep analysis of the characteristics of the unlabeled target domain. These provide relevant information not only for the selection of the best learning algorithm [3], but also for the selection of domains with a feature distribution similar to the target domain. In this way, the use of metalearning techniques by Shou-Shan et al. [14] can be enriched by knowledge such as the performance distribution of dissimilar domains to the target data.

Table 3. Scores obtained with SO

	Books	**Movies**	**Music**	**Phones**
Books		65.31%	59.18%	59.18%
Movies	72.00%		62.00%	56.00%
Music	68.00%	74.00%		52.00%
Phones	77.55%	51.02%	65.31%	

5 Conclusions and Future Work

In this paper, we approached the adaptation problem in sentiment classification by determining the subset of domains with a feature distribution similar to the unlabeled target domain. In particular, our approach intends to show how the estimation of similarity distribution among different domains provides useful information for the classification of unlabeled reviews. Our framework is based on the definition of domains' profiles as well as similarity metrics used to identify unseen reviews. The results show how a feature distribution based on the semantic orientation of available datasets is a plausible option for the classification of unlabeled reviews.

As part of our future work, we intend to explore the behavior of our method with different datasets. In particular, we are interested in the use of the dataset collected by Blitzer et al. [4]. This dataset is basically a collection of product reviews from four domains: books, DVDs, electronics, and kitchen appliances. Since each domain contains 1,000 positive and 1,000 negative reviews, we think this collection is worth our attention.

Finally, we also intend to deep in the exploration of a meta-level system to identify the most similar datasets for a given input dataset [3]. The purpose is to understand the connection between learning algorithms and the characteristics of the data under analysis.

References

1. Aue, A., Gamon, M.: Customizing sentiment classifiers to new domains: A case study. In: Proceedings of Recent Advances in Natural Language Processing (RANLP), Borovets, Bulgaria (2005)
2. Liu, B.: Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, 1st edn. Springer (2007)
3. Brazdil, P., Giraud-Carrier, C., Soares, C., Vilalta, R.: Metalearning Applications to Data Mining, 1st edn. Springer (2009)
4. Blitzer, J., Dredze, M., Pereira, F.: Biographies, Bollywood, Boom-boxes and Blenders: Domain adaptation for sentiment classification. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL), Prague, Czech Republic (2007)
5. Chesley, P., Vincent, B., Xu, L., Srihari, R.: Using verbs and adjectives to automatically classify blog sentiment. In: AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW), pp. 74–80 (2006)
6. Escalante, H.J., Montes, M., Solorio, T.: A Weighted Profile Intersection Measure for Profile-based Authorship Attribution. In: Proceedings of 10th Mexican International Conference on Artificial Intelligence (MICAI), Puebla, Mexico (2011)
7. Esuli, A., Sebastiani, F.: SentiWordNet: a publicly available lexical resource for opinion mining. In: Proceedings of Fifth International Conference on Language Resources and Evaluation (LREC), Genoa, Italy, vol. 6 (2006)
8. Heck, T.: A Comparison of Different User-Similarity Measures as Basis for Research and Scientific Cooperation. In: International Conference on Information Science and Social Media, bo/Turku, Finland (2011)

9. Keefe, T., Koprinska, I.: Feature Selection and Weighting Methods in Sentiment Analysis. In: Proceedings of the 14th Australasian Document Computing Symposium, Sydney, Australia (2009)
10. Keselj, V., Peng, F., Cercone, N., Thomas, C.: N-gram-based author profiles for authorship attribution. In: Proceedings of the Pacific Association for Computational Linguistics, Halifax, Canada, pp. 255–264 (2003)
11. Liu, B.: Sentiment Analysis and Subjectivity. In: Indurkhya, N., Damerau, F.J. (eds.) Handbook of Natural Language Processing, pp. 627–665. Chapman and Hall/CRC (2010)
12. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment Classification Using Machine Learning Techniques. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Philadelphia, USA, pp. 79–86 (2002)
13. Pang, B., Lee, L.: 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), Barcelona, Spain, pp. 21–26 (2004)
14. Shou-Shan, L., Chu-Ren, H., Cheng-Qing, Z.: Multi-Domain Sentiment Classification with Classifier Combination. Journal of Computer Science and Technology 26(1), 25–33 (2011)
15. Steck, J.B.: Netpix: A Method of Feature Selection Leading to Accurate Sentiment-Based Classification Models. Master thesis, Central Connecticut State University (2005)
16. Stamatatos, E.: A survey of modern authorship attribution methods. Journal of the American Society for Information Science and Technology 60(3), 538–556 (2009)
17. Taboada, M., Anthony, C., Voll, K.: Creating Semantic Orientation Dictionaries. In: Proceedings of Fifth International Conference on Language Resources and Evaluation (LREC), Genoa, Italy, pp. 427–432 (2006)
18. Toutanova, K., Klein, D., Manning, C., Singer, Y.: Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In: Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL), Edmonton, Canada, pp. 252–259 (2003)
19. Turney, P.: Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, USA, pp. 417–424 (2002)
20. Yang, H., Si, L., Callan, J.: Knowledge transfer and opinion detection in the TREC2006 blog track. In: Proceedings of the Fifteenth Text REtrieval Conference, Gaithersburg, MD (2006)