Discriminative Bi-Term Topic Model for Headline-Based Social News Clustering

Yunqing Xia
School of Computer Science
Tsinghua University
yqxia@tsinghua.edu.cn

Nan Tang
School of Computer Science
Carnegie Mellon University
gracytang92@gmail.com

Amir Hussain
School of Natural Sciences
University of Stirling
ahu@cs.stir.ac.uk

Erik Cambria
School of Computer Engineering
Nanyang Technological University
cambria@ntu.edu.sg

Abstract

Social news are becoming increasingly popular. News organizations and popular journalists are starting to use social media more and more heavily for broadcasting news. The major challenge in social news clustering lies in the fact that textual content is only a headline, which is much shorter than the fulltext. Previous works showed that the bi-term topic model (BTM) is effective in modeling short text such as tweets. However, the drawback is that all non-stop terms are considered equally in forming the bi-terms. In this paper, a discriminative bi-term topic model (d-BTM) is presented, which tries to exclude less indicative bi-terms by discriminating topical terms from general and document-specific ones. Experiments on TDT4 and Reuter-21578 show that using merely headlines, the d-BTM model is able to induce latent topics that are nearly as good as that are generated by LDA using news fulltext as evidence. The major contribution of this work lies in the empirical study on the reliability of topic modeling using merely news headlines.

Introduction

In recent years, news organizations and popular journalists started to use social media primarily to broadcast news articles. Another study by The American Press Institute indicates that 44 percent Americans discover news on social networks (see *How Americans get their news* at http://www.americanpressinstitute.org/). Social news usually contain a short headline and a link to the news fulltext in the source website (Figure 1).

Due to length limit, only title and/or headline can be visible to the followers. Note that the followers do not always click the links in the social news articles to view the story. Only interesting topics make them do so. Thus, it is demanded that topics should be presented based on news headlines before people view the full text.

As a successful solution to news management, the text clustering methods now encounter a major challenge in handling the social news. However, dealing with the social news, the state-of-the-art text clustering methods encounter serious sparse data problem.

That is, the available textual content is very short. Experimental results show that with very short textual content, news clustering systems suffer significant quality loss from the sparse data problem. In Wikipedia, *headline* is defined as *the text indicating the nature of the article below it*. As its purpose is *to quickly and briefly draw attention to the story*, headline is usually informative and accurate in professional writing. Therefore it is professional to present only headlines in the social news. With this observation, we ambitiously assume in this work that headline alone can yield reasonable clustering quality if proper models are adopted. In this work, we concentrate on designing topic models for the headline-based social news clustering.

We start from the *bi-term topic model* proposed in (Yan et al. 2013), which is proved effective in handling short text. It simply considers all non-stop terms equally in forming the bi-terms. However, we notice that in news headlines, non-stop terms are not equally indicative in presenting the topics. For the headline in Figure 1, term *battery* is topical term which plays a major role in presenting the topic. As a comparison, term *scientist* is a rather general term and it contributes little to the topic. In fact, terms in news headlines can be categorized into topical, general and document-specific. We argue that the category information is rather important in selecting appropriate terms for topic modeling.

Enlightened by this observation, we propose the discriminative bi-term topic model (d-BTM) which excludes less indicative bi-terms by discriminating topical terms from general and document-specific ones. In this work, we make use document ratio and design simple rules to achieve the classification. We realize there is no gold-standard dataset for social news clustering. However, we have a few datasets for regular news articles. We thus extract the first sentence in a news article to simulate textual content of a social news.

Experimental results on three news article datasets: *TDT*41, *TDT*42 and *Reuters*20, indicate that the proposed headline-based discriminative bi-term topic model is as effective as LDA which uses fulltext as news clustering evidence.

A major contribution of this work is the empirical study conducted on the reliability of topic modeling using only headlines. The rest of the paper is organized as follows: we summarize related work in the next section; then, we present the *d*-BTM model; next, experiments are presented in an evaluation section; finally, we conclude the paper.

Figure 1: An example of BBC news on twitter

Related Work

Existing approaches to sentiment analysis can be grouped into four main categories: keyword spotting, lexical affinity, statistical methods, and concept-based techniques. Keyword spotting is the most naïve approach and probably also the most popular because of its accessibility and economy. Text is classified into affect categories based on the presence of fairly unambiguous affect words like 'happy', 'sad', 'afraid', and 'bored'. Elliott's Affective Reasoner (Elliott 1992), for example, watches for 198 affect keywords, e.g., 'distressed' and 'enraged', plus affect intensity modifiers, e.g., 'extremely', 'somewhat', and 'mildly'.

Lexical affinity is slightly more sophisticated than keyword spotting as, rather than simply detecting obvious affect words, it assigns arbitrary words a probabilistic 'affinity' for a particular emotion. For example, 'accident' might be assigned a 75% probability of being indicating a negative affect, as in 'car accident' or 'hurt by accident'. These probabilities are usually trained from linguistic corpora (Wilson, Wiebe, and Hoffmann 2005; Stevenson, Mikels, and James 2007; Somasundaran, Wiebe, and Ruppenhofer 2008).

Statistical methods, such as support vector machines (Vapnik and Kotz 2006), deep learning (Lee et al. 2011) and extreme learning machines (Cambria, Huang, and et al. 2013), have been popular for affect classification of texts and have been used by researchers on projects such as Pang's movie review classifier (Pang, Lee, and Vaithyanathan 2002) and many others (Hu and Liu 2004; Glorot, Bordes, and Bengio 2011; Socher et al. 2013; Lau, Xia, and Ye 2014; Cambria et al. 2015b).

By feeding a machine learning algorithm a large training corpus of affectively annotated texts, it is possible for the system to not only learn the affective valence of affect keywords (as in the keyword spotting approach), but also to take into account the valence of other arbitrary keywords (like lexical affinity) and word co-occurrence frequencies.

However, statistical methods are generally semantically weak, i.e., lexical or co-occurrence elements in a statistical model have little predictive value individually. As a result, statistical text classifiers only work with acceptable accuracy when given a sufficiently large text input. So, while these methods may be able to affectively classify user's text on the page- or paragraph-level, they do not work well on smaller text units such as sentences or clauses.

Concept-based approaches, in turn, focus on a semantic analysis of text by means of semantic networks (Poria et al. 2014), web ontologies (Gangemi, Presutti, and Reforgiato 2014), or semantic multidimensional scaling (Cambria et al. 2015a). Rather than working only at data- or syntactic-level, concept-based approaches take into account additional knowledge, e.g., the semantics and sentics (i.e., denotative and connotative information) associated with natural language opinions. By relying on large semantic knowledge bases, such approaches step away from blind use of keywords and word co-occurrence count, but rather rely on the implicit meaning/features associated with natural language concepts. Unlike purely syntactical techniques, in fact, concept-based approaches are able to detect also sentiments that are expressed in a subtle manner, e.g., through the analysis of concepts that do not explicitly convey any emotion, but which are implicitly linked to other concepts that do so.

Clustering algorithm became a major solution to topic analysis on news articles since TDT1 initiated the task of topic detection and tracking (Allan, Carbonell, and et al. 1998). In the past fifteen years, a great majority of research efforts have been made on topic modeling which seeks to represent news articles using fulltext. An early attempt is probabilistic Latent Semantic Indexing (Hofmann 1999) followed by Latent Dirichlet allocation (LDA) (Blei, Ng, and Jordan 2003). In (Tang et al. 2014b), word sense is further incorporated in LDA to make the model more flexible to meanings. In (Tang et al. 2014a), the word sense based LDA is applied to cross-lingual text clustering. All the previous work makes use of news full text in clustering. Our work is different because we attempt to achieve the same goal using merely headline, which is much shorter in length.

Though little work is reported on headline-based news clustering, we notice that some related work on shorter text clustering is also enlightening. For example, a Web-based kernel function is proposed in (Sahami and Heilman 2006) to measure similarity of short text snippets returned by search engines. In (Jin et al. 2011), a dual Latent Dirichlet allocation model is proposed to jointly learn two sets of topics on short and long texts and couples the topic parameters to cope with the potential inconsistencies between data sets. In (Yan et al. 2013), a bi-term topic model is proposed to deal with short social media texts.

In this work, we attempt to tackle the problem of the

headline-based news clustering using the bi-term topic model while the difference between our model and the original one lies in that we discriminate different types terms in forming the bi-terms.

Model

Motivation

The bi-term topic model is proved effective in clustering both tweets and normal text (Yan et al. 2013). However, it simply considers all non-stop terms equally in forming the bi-terms, which is basically not true in news articles. Observation on the news article dataset indicates that terms appearing the news articles can be classified into three categories:

- **Topical terms** (T): the T terms are representative and discriminative to some topics in the dataset. For example, *battery* in Figure 1 is a topical term, which helps to represent or discriminate this news.
- **General terms** (G): the G terms are general to many topics in the dataset. For example, *scientist* in Figure 1 is a general term. It helps to give background while they must be associated with topical terms to present a special topic.
- **Document-specific terms** (D): the D terms appear in only one document in the dataset. Terms of this kind are usually unique to a document. However, they usually make little contribution to topic because their document frequency is rather low (i.e., 1).

In news articles, the T terms are most indicative while the G terms are least. Considering the bi-term topic model, we have 6 types of bi-terms: T-T, T-G, T-D, G-G, G-D and D-D. In fact, according to definitions of G term and D term, the G-G, G-D and D-D bi-terms are obviously not indicative in topic modeling. We exclude the three types of bi-terms, and finally we have the T-T, T-G and T-D bi-terms.

We argue that the bi-term topic model can be improved by discriminating the type of these bi-terms. Thus, we propose the discriminative bi-term topic model (d-BTM) to implement such an idea. For presentation convenience, we first brief the bi-term topic model in the following section.

Bi-term Topic Model

Details of the bi-term topic model is given in (Yan et al. 2013). We only give the generation process below.

1. For each topic z:
   (a) draw a topical word distribution $\phi_z \sim Dir(\beta)$
2. Draw a topic distribution $\theta \sim Dir(\alpha)$ for the whole collection
3. For each bi-term b in the bi-term set B:
   (a) draw a topic assignment $z \sim Multi(\theta)$
   (b) draw two words: $w_i, w_j \sim Multi(\phi_z)$

Our revision on the bi-term topic model happens in step 3(b), in which two words w_i, w_j are drawn to form the bi-term $w_i - w_j$.

In this work, we first assign each term a category (i.e., T, G or D). Then we select terms in appropriate categories to form the bi-term. To be specific, the following three types of bi-terms are selected: T-T, T-G and T-D. We will conduct an experiment to evaluate which combination is more effective.

Term Classification

A straightforward question is how the terms are assigned the aforementioned three categories. In this work, we adopt simple rules based on corpus statistics to achieve this goal. We first recognize the G terms then G terms, and consider the remaining terms as T terms.

G terms

According to the definition, the general terms in fact provide the function of presenting background of topics in the dataset. They should appear in many documents which are assigned different topics. We thus define the document ratio to reflect how general a term is. To be formal, for a specific term t, its document ratio $r^D(t)$ is calculated as follows,

$$r^D(t) = \frac{df(t)}{|D|}, \tag{1}$$

where $df(t)$ represents number of documents in which term t is mentioned, and $|D|$ total number of documents in the dataset.

Intuitively, a higher document ratio indicates that the term is more general. Thus we may set a threshold and consider terms with bigger document ratio values as general terms. The threshold is rather empirical. There is in fact no theoretical proof saying which threshold is better. However, we may start defining the threshold from counting number of documents in every topic.

Let T_i denote the i-th topic, $c^D(T_i)$ denote number of documents in topic T_i. To make sure that a term t appears in at leat two topics (to be general), we should require that document frequency $df(t)$ of term t is bigger than $c^D(T^*)$ where T^* denotes the biggest topic in the dataset.

Thus, we calculate the threshold value for every dataset as follows,

$$r^D_{\mathrm{T}} = \max_i \frac{c^D(T_i)}{|D|}. \tag{2}$$

For example, the TDT41 dataset contains 657 documents, thus $|D|$=657. As the biggest topic in TDT41 contains 154 documents, the threshold value r^D_{T} for this dataset is $154/657 \approx 0.234$. For the specific term *research*, its document ratio $df(t) = 186/657 \approx 0.283$. As $0.283 > 0.234$, term *research* is deemed as a general term.

D terms

We simply consider terms that appear in only one document as document-specific terms.

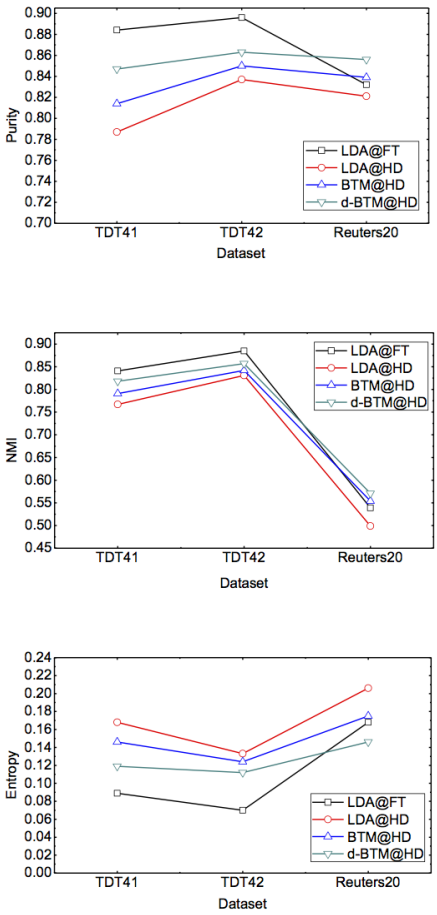

Figure 2: Results of the topic models.

Evaluation

Setup

Dataset

To the best of our knowledge, there is no corpus available for topic annotations of social news. Fortunately, we have a few datasets for news fulltext with topic annotations. We intend to make use of the fulltext annotations to evaluate the headline-based news clustering methods. Extracting headline from new fulltext is another research topic.

For simplicity, we use the first first sentence in each news article as headline. This is true for most news reports which follow the writing style of presenting the most important information in the first sentence.

- *TDT*41: TDT2002 part in TDT4 dataset (Kong and Graff 2005). It involves 38 topics within 1,270 news articles.
- *TDT*42: TDT2003 part in TDT4 dataset (Kong and Graff 2005). It covers 33 topics in 617 news articles.
- *Reuters*20: Part of Reuters-21578 (Lewis 1997). It contains 9,101 documents extracted from the 20 biggest categories in Reuters-21578.

Evaluation Metrics

We evaluate the proposed *d*-BTM model with the following evaluation metrics adopted in (Yan et al. 2013):

- *Purity*: is used to determine quality of clusters by measuring the extent to which each cluster contains documents primarily from one category. When all the documents in each cluster are assigned with the same class, purity in the experiment is highest with value of 1. On the contrary, purity is close to 0 in bad clustering.
- *NMI*: is to identify the trade-off between quality of the clusters against the number of clusters. Note NMI=1 indicates a perfect match between system output and gold standard, while 0 indicates a random clustering random with respect to class membership.
- *Entropy*: is to measure distribution of the documents on various categories. A lower entropy usually indicates a better clustering performance.

Workflow for News Clustering

We adopt a general workflow for news clustering which is accomplished in four steps. First, latent topics are extracted from news collection using topic models. Second, each news article is represented with a topic vector in which the latent topics are considered features and the probability values of terms (or bi-terms) in the topics are considered as weights. Third, we calculate vector similarity using cosine equation. At last, we run Bisecting K-Means to group the news articles due to its excellent performance in text clustering (Steinbach, Karypis, and Kumar 2000).

The Topic Models

In this experiment, we aim at comparison between the topic models in performing the task of news clustering using various length of news articles. The following topic models are compared in this experiment:

- **LDA@FT**: Modeling news articles with LDA (Blei, Ng, and Jordan 2003) using news article fulltexts.
- **LDA@HD**: Modeling news articles with LDA (Blei, Ng, and Jordan 2003) using news article headlines.
- **BTM@HD**: Modeling news articles with BTM (Yan et al. 2013) using news article headlines.
- ***d*-BTM@HD**: Modeling news articles with the our proposed *d*-BTM using news article headline considering merely topical (T) terms in forming the bi-terms.

The parameters involved in our evaluation are LDA-related. Following the previous work, we set α=0.1, β=0.01, and K to be number of the topics in the evaluation dataset.

We present experimental results on *purity*, *NMI* and *entropy* in news clustering in Figure 2.

Discussion

An important observations are made on Figure 2. When the fulltexts are replaced by the headlines, quality of clustering results with LDA drops significantly (e.g., 0.078 on average on *purity*) on three metrics on two TDT datasets

while slightly(e.g., 0.011 on *purity*) on Reuters20 dataset. Also, with headlines, BTM improves the quality slightly (e.g., 0.019 on average on *purity*). On TDT1 and TDT2, the proposed *d*-BTM model further improves the quality (e.g., by 0.04 on average on *purity*) to a level that is much closer to the fulltext-based LDA. It is even surprising that on Reuters20, our headline-based *d*-BTM model outperforms fulltext-based LDA slightly (e.g., 0.024 on *purity*).

This indicates that headlines in TDT datasets are more sensitive to high-quality headlines. Study on news articles in Reuters20 indicates that headlines in the Reuters news are finely compiled, which is self-contained in about 30 words. As a comparison, headlines in TDT news are relatively short and partial on topic presentation. This implies that with finely-compiled headlines, *d*-BTM is able to yield reliable clusters of news articles. This work thus provides empirical evidence that news article clustering can be satisfactorily achieved with much shorter content.

Term discrimination

We aim at observing contribution of term discrimination in this experiment. To achieve this goal, we implement the following implementations of *d*-BTM model with different types of bi-terms.

- **BTM**: The original BTM modeling (Yan et al. 2013) using all terms in forming the bi-terms.
- ***d*-BTM**(-D): our *d*-BTM modeling excluding document-specific terms in forming the bi-terms.
- ***d*-BTM**(-G): our *d*-BTM modeling excluding general terms in forming the bi-terms.
- ***d*-BTM**(-GD): our *d*-BTM modeling excluding document-specific terms and general terms in forming the bi-terms (i.e., *d*-BTM@HD in the last experiment).

In this experiment, we perform *d*-BTM topic modeling on headlines. Experimental results on entropy, *purity* and *NMI* in news clustering are presented in Figure 3.

Discussion

Seen from Figure 3, our *d*-BTM model receives consistent performance gain (e.g., 0.021 on average on *purity*) by excluding general terms and/or document-specific terms in forming the bi-terms. This indicates that removing document-specific terms and general terms is helpful in *d*-BTM topic modeling.

Comparing the three *d*-BTM implementations, we find they perform slightly different on the three datasets. On TDT41 and Reuters20, *d*-BTM(-GD) performs best. However, on TDT42, *d*-BTM(-G) performs best. Looking into news articles in the three datasets, we find the difference is caused by nature of the news. Fewer general terms are detected in TDT42 than that in TDT41 and Reuters20, due to a bigger document ratio threshold.

Thus removing the general terms makes the latent topics more precise in TDT41. Meanwhile, the headlines in TDT42 contain more document-specific terms than that in TDT41 and Reuters20. Thus the latent topics are made less precise by removing the document-specific terms in TDT41.

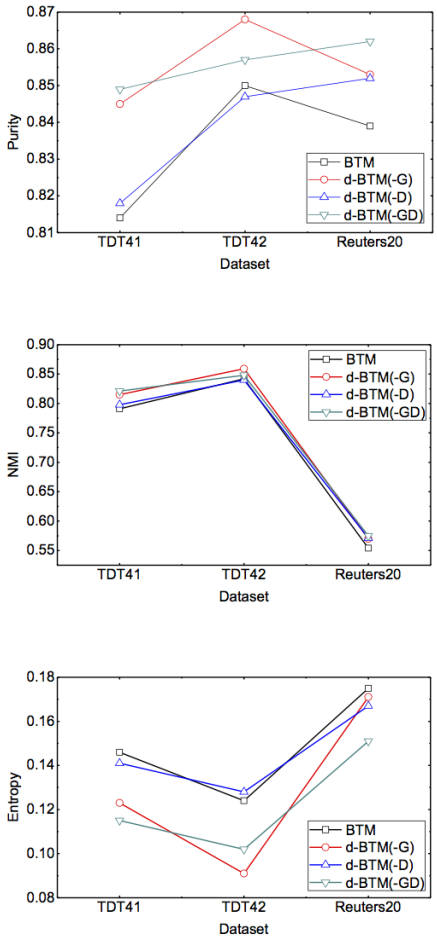

Figure 3: Results of the *d*-BTM implementations.

Conclusion and Future Work

To feed the demand of social news clustering, we propose the discriminative bi-term topic model (*d*-BTM) in this work to achieve the clustering goal based on headlines. Compared to the original BTM model which uses all terms in forming the bi-terms, *d*-BTM attempts to exclude terms that are less indicative to topic modeling by classifying term to be topical, document-specific and general. Contribution of this work lies mainly in the empirical study on the reliability of topic modeling using merely headlines. Experimental results show that the proposed *d*-BTM model with headlines can yield latent topics as good as those induced by LDA using news fulltexts. This guarantees the reliability of social news clustering methods which use merely headlines as evidence.

This work is still preliminary. The following future work is planned. We will design better algorithm for term discrimination, especially in general term detection. We will also conduct experiments to compare the proposed model against the baseline models regarding computing time.

Acknowledgments

This work is partially supported by the National Science Foundation of China (61272233). We thank the anonymous reviewers for their insightful comments.

References

Allan, J.; Carbonell, J.; and et al. 1998. Topic detection and tracking pilot study final report. In *Proc. of the DARPA Broadcast News Transcription and Understanding Workshop*, 194–218.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.* 3:993–1022.

Cambria, E.; Fu, J.; Bisio, F.; and Poria, S. 2015a. AffectiveSpace 2: Enabling affective intuition for concept-level sentiment analysis. In *AAAI*.

Cambria, E.; Gastaldo, P.; Bisio, F.; and Zunino, R. 2015b. An ELM-based model for affective analogical reasoning. *Neurocomputing* 149:443–455.

Cambria, E.; Huang, G.-B.; and et al. 2013. Extreme learning machines. *IEEE Intelligent Systems* 28(6):30–59.

Elliott, C. D. 1992. *The Affective Reasoner: A Process Model of Emotions in a Multi-Agent System*. Ph.D. Dissertation, Northwestern University, Evanston.

Gangemi, A.; Presutti, V.; and Reforgiato, D. 2014. Frame-based detection of opinion holders and topics: a model and a tool. *IEEE Computational Intelligence Magazine* 9(1):20–30.

Glorot, X.; Bordes, A.; and Bengio, Y. 2011. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *ICML*.

Hofmann, T. 1999. Probabilistic latent semantic indexing. In *Proc. of SIGIR'1999*, 50–57. New York, NY, USA: ACM.

Hu, M., and Liu, B. 2004. Mining and summarizing customer reviews. In *KDD*.

Jin, O.; Liu, N. N.; Zhao, K.; Yu, Y.; and Yang, Q. 2011. Transferring topical knowledge from auxiliary long texts for short text clustering. In *Proc. of CIKM'2011*, 775–784. New York, NY, USA: ACM.

Kong, J., and Graff, D. 2005. Tdt4 multilingual broadcast news speech corpus. *Linguistic Data Consortium, http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp*.

Lau, R.; Xia, Y.; and Ye, Y. 2014. A probabilistic generative model for mining cybercriminal networks from online social media. *IEEE Computational Intelligence Magazine* 9(1):31–43.

Lee, H.; Grosse, R.; Ranganath, R.; and Ng, A. Y. 2011. Unsupervised learning of hierarchical representations with convolutional deep belief networks. *Communications of the ACM* 54(10):95–103.

Lewis, D. D. 1997. Reuters-21578 text categorization test collection, distribution 1.0. *http://www.research.att.com/˜lewis/reuters21578.html*.

Pang, B.; Lee, L.; and Vaithyanathan, S. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *EMNLP*, 79–86.

Poria, S.; Gelbukh, A.; Cambria, E.; Hussain, A.; and Huang, G.-B. 2014. EmoSenticSpace: A novel framework for affective common-sense reasoning. *Knowledge-Based Systems* 69:108–123.

Sahami, M., and Heilman, T. D. 2006. A web-based kernel function for measuring the similarity of short text snippets. In *WWW*, 377–386. New York, NY, USA: ACM.

Socher, R.; Perelygin, A.; Wu, J. Y.; Chuang, J.; Manning, C. D.; Ng, A. Y.; and Potts, C. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*.

Somasundaran, S.; Wiebe, J.; and Ruppenhofer, J. 2008. Discourse level opinion interpretation. In *COLING*, 801–808.

Steinbach, M.; Karypis, G.; and Kumar, V. 2000. A comparison of document clustering techniques. In *Proc. of KDD Workshop on Text Mining*.

Stevenson, R.; Mikels, J.; and James, T. 2007. Characterization of the affective norms for english words by discrete emotional categories. *Behavior Research Methods* 39:1020–1024.

Tang, G.; Xia, Y.; Cambria, E.; Jin, P.; and Zheng, T. 2014a. Document representation with statistical word senses in cross-lingual document clustering. *International Journal of Pattern Recognition and Artificial Intelligence*.

Tang, G.; Xia, Y.; Sun, J.; Zhang, M. Z.; and Zheng, T. F. 2014b. Statistical word sense aware topic models. *Soft Computing*.

Vapnik, V., and Kotz, S. 2006. *Estimation of Dependences Based on Empirical Data*. Springer.

Wilson, T.; Wiebe, J.; and Hoffmann, P. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *HLT/EMNLP*, 347–354.

Yan, X.; Guo, J.; Lan, Y.; and Cheng, X. 2013. A biterm topic model for short texts. In *Proc. of WWW'2013*, 1445–1456.