Mining Sentiment Terminology Through Time

Hadi Amiri
Department of Computer Science
NUS Graduate School for Interactive Science and Engineering
National University of Singapore, 117417
hadi@comp.nus.edu.sg

Tat-Seng Chua
Department of Computer Science
NUS Graduate School for Interactive Science and Engineering
National University of Singapore, 117417
chuats@comp.nus.edu.sg

ABSTRACT

The correspondence between sentiment terminology and the active language used for expressing opinions is a crucial prerequisite for effective sentiment analysis. Mining sentiment terminology includes the detection of new opinion words as well as inferring their polarities. In this paper, we first propose a novel approach based on the *interchangeability* characteristic of words to detect new opinion words through time. We then show that the current non-time-based polarity inference approaches may assign opposite polarity to the same opinion word at different times. To tackle this issue, we consider the polarity scores computed at different times as polarity *evidences* (with the possibility of flawed evidences) and combine them to compute a globally correct polarity score for each opinion word. The experiments show that our approach is effective both in terms of the quality of the discovered new opinion words as well as its ability in inferring their polarities through time. Furthermore, we show the application of mining sentiment terminology through time in the sentiment classification (SC) task. The experiments show that mining more recent new opinion words leads to greater improvement in the performance of SC. To the best of our knowledge, this is the first work that investigates "time" as an important factor in mining sentiment terminology.

Categories and Subject Descriptors

I.2.7 [**Natural Language Processing**]: Text Analysis;; H.3.1 [**Content Analysis and Indexing**]: Linguistic processing

General Terms

Algorithms, Experimentation.

Keywords

Opinion Word Mining, Temporal Opinion Lexicon, Sentiment Orientation, Word Polarity

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.

1. INTRODUCTION

Current sentiment lexicons, contain most of the common opinion words (like *excellent* and *amazing*), but they miss emerging sentiment words like slang and urban opinion words (such as *delish* and *cozy*) as well as the misspelled common opinion words (like *excellance* and *recomend*). However, such words are commonly used in user generated contents (UGCs) to express opinions. As such, there is a need to develop techniques to address the big gap between the current sentiment lexicons and the emerging opinion words. This includes the detection of new opinion words as well as inferring their polarity through time.

Previous research have mainly utilized synthetic and co-occurrence patterns to mine new opinion words [9, 1]. In this research, we show that "time" is another important factor for mining sentiment words in the sense that: (a) new opinion words emerge at different times as UGC is growing, (b) the current methods based on synthetic and co-occurrence patterns often estimate different polarity for opinion words at different times, and (c) though rarely happen, opinion words may change their sentiment orientation through time. For example, the term "*awesome*" meant "*terrifying*" in the past, but nowadays it means "*amazing*".

Figure 1 illustrates the polarity scores of several new opinion words estimated by the popular non-time-based Turney and Littman's (2003) method at different times (the time granularity is six months). As Figure 1 shows, for each word, the polarity scores are often wrongly estimated at different times and vary through time. This is because of the varying co-occurrence patterns observed at different times.

To tackle the above challenges, we first utilize the *interchangeability* characteristic of words to detect new opinion words through time. We then propose a novel polarity inference technique to infer *time accumulated* polarity scores for the new opinion words. We consider the polarity scores obtained at different times as polarity *evidences* and combine them to compute the time accumulated polarity scores. For this purpose, we use the Dempster-Shafer combination theory [2, 4] which is known to be strong with respect to flawed evidences. We show that this consideration leads to more accurate polarity inference.

To summarize, the contributions of this paper are as follows: (a) we propose a time-aware approach to mine sentiment terminology, (b) we show that polarity accumulation through time result in a more accurate polarity inference than the polarity obtained using the non-time-based methods at different times, and (c) we show that mining more

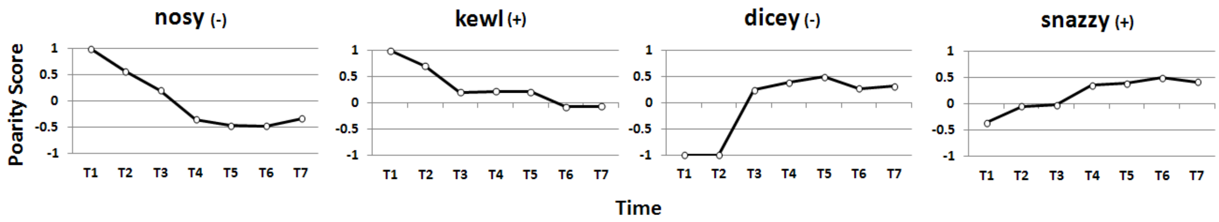

Figure 1: Polarity scores computed by the Turney and Littman's (2003) method at different times.

recent new opinion words leads to greater improvement in the performance of sentiment classification.

The average performance of our method in detecting new opinion words is 68.76%. Furthermore, our approach significantly outperforms the state-of-the-art non-time-based polarity inference method by 5.8% on average in F1 score. In addition, experiments on sentiment classification (SC) of reviews show that the new opinion words significantly improve the F1 performance over the baseline through time.

The rest of this paper is organized as follows: Section 2 elaborates our approach for mining sentiment terminology through time. Section 3 reports the experimental settings and results. Section 4 surveys the related work and, finally, Section 5 concludes the paper and discusses future directions[1].

2. MINING NEW OPINION WORDS

In this section we present a context-aware approach to mine new opinion words through time.

2.1 Interchangeability

We propose to find the interchangeable words that are distributionally similar with seeds (words with already known polarity) and consider them as candidate new opinion words. We define the interchangeability between two words as follows:

Definition 1: Two words are interchangeable, if they have:

1. low co-occurrence,
2. high overlap in their left neighboring words, and
3. high overlap in their right neighboring words.

Due to the intuitive definition of interchangeability, the co-occurrence between two interchangeable words is expected to be low. For example, since "*suggest*" and "*recommend*" are interchangeable, we usually use one of them in a sentence to give a suggestion. Furthermore, we here separately deal with the left and right neighboring words to discard the effect of the words that occur on the opposite sides of the target words in measuring their interchangeability.

To find interchangeable words with seeds, we assume that the time-span T_i includes all the reviews written in the time interval $[t_{i-1}, t_i]$. Let T_i, $i \leq j$, be the *source* time-span and T_j be the *target* time-span. The words of T_j that are interchangeable with at least one seed of T_i are candidate new opinion words. Given two words a^i and b^i from the same time-span T_i, we first define the side-oriented PMI between them as follows:

$$
\begin{aligned}
&PMI^l(a^i, b^i) = \\
&\qquad \log\left(\frac{Count^i(a^i \text{ occur on left side of } b^i)M^i}{Count^i(a^i)Count^i(b^i)}\right) \\
&PMI^r(a^i, b^i) = \\
&\qquad \log\left(\frac{Count^i(a^i \text{ occur on right side of } b^i)M^i}{Count^i(a^i)Count^i(b^i)}\right)
\end{aligned} \tag{1}
$$

where $Count^i(x)$ is the number of sentences that contain x at time-span T_i, and M^i is the number of sentences at T_i.

In addition, given the word a^i from the time-span T_i, we refer to its left (right) significant neighboring words (SNWs) as the words of T_i that (1) occur on the left (right) side of a^i, and (2) have positive PMI^l (PMI^r) values with respect to a^i. For each word, we only consider its top z left (right) SNWs that have the highest PMI^l (PMI^r) values with respect to the word.

Let v^i be a *seed* word from T_i, $i \leq j$, and w^j be a target word from T_j. We define $S^l_{v^i w^j}$ and $S^r_{v^i w^j}$ as the common left and right SNWs of v^i and w^j respectively and compute the *context similarity* between the two words as follows:

$$
\begin{aligned}
Sim(v^i, w^j) = \frac{1}{z} &\sum_{O\in\{l,r\}} \sum_{u\in S^O_{v^i w^j}} \\
&[(PMI^O(v^i, u^i))^\zeta + (PMI^O(u^j, w^j))^\zeta]
\end{aligned} \tag{2}
$$

where O indicates left or right, u is a common (left or right) SNW of both v^i and w^j, and ζ is a constant. Equation (2) computes the similarity between two words by aggregating the PMI values of their common left and right SNWs. It assigns high similarity scores to the words that either (1) frequently co-occur, or (2) rarely co-occur but have high semantic association, such as "*recommend*" and "*suggest*". According to Definition 1, we are only interested in the latter case, so, we discard the words that frequently co-occur. For this purpose, we use the side-oriented PMI as the measure of co-occurrence and compute the interchangeability score between two eords as follows:

$$
\begin{aligned}
&Int(v^i, w^j) = \\
&\frac{Sim(v^i, w^j)}{c + \sum_{O\in\{l,r\}} PMI^O(v^i, w^i) + PMI^O(v^j, w^j)}
\end{aligned} \tag{3}
$$

where c is a small constant. We construct an interchangeability pool, $\mathcal{P}_{ij}$, for each source-target time-pair, (T_i,T_j) $\forall i \leq j$, as follows:

$$
\mathcal{P}_{ij} = \{w^j_1, w^j_2, ...\} \tag{4}
$$

where each $w^j_k \in \mathcal{P}_{ij}$ is a candidate new opinion word of T_j that is interchangeable with at least one seed of T_i.

[1]This research is done as part of NExT projects. NExT Search Center is supported by the Singapore National Research Foundation & Interactive Digital Media R&D Program Office, MDA under research grant (WBS:R-252-300-001-490).

2.2 Non-Time-Based Polarity Inference

We utilize a non-time-based approach to first assign a polarity score to each candidate new opinion word w_k^j that appears in an interchangeability pool. In particular, for each w_k^j, we use all the reviews up to time T_j to compute the polarity score of w_k^j obtained at time T_j. This will be considered as a polarity evidence for the word w in the future.

We use the optimization framework proposed in [1] to compute the polarity scores at different times. In this framework, seeds and candidate opinion words are respectively treated as labeled and unlabeled data and modeled in a so-called *polarity graph*. Edges are weighted by a function of the co-occurrence between their nodes. This framework optimized the prediction for unlabeled nodes by assigning similar polarity scores to nodes that are connected with heavy weighted edges in the polarity graph.

In our setting, we consider each candidate opinion word $w_k^j \in \mathcal{P}_{ij}$ as an unlabeled node that, at the end of this process, will be assigned a polarity score $f(w_k^j)$ by the optimization framework. We refer to this value as a polarity evidence for the word w obtained at T_j and show it by $Pol(w_k^j)$.

2.3 Time Accumulated Polarity Inference

As we elaborated before, non-time-based polarity inference methods may assign opposite polarity scores to a given opinion word at different times. This is mainly because such methods rely on the noisy co-occurrence patterns. To tackle this issue, we compute a time accumulated polarity score for each candidate opinion word using its polarity scores obtained at different times. For this purpose, we utilized the Dempster-Shafer (DS) combination theory as it is strong with respect to the flawed evidences.

In the DS theory [2, 4] there exist a set of mutually exclusive alternatives which is called the *frame of discriminant* Θ. For example, for opinion words, Θ can be defined as follows:

$$\Theta = \{positive, negative\} \tag{5}$$

The DS theory assigns a *belief* value to each element of the power set of Θ. Formally, the function $m : 2^\Theta \to [0, 1]$ is called *basic probability assignment* (BPA), if it has the following properties:

$$m(\phi) = 0, \sum_{A \in 2^\Theta} m(A) = 1 \tag{6}$$

where $m(A)$ is the belief value that the proposition $A \in 2^\Theta$ is true for an observation (a word here). Obviously, the belief values of the power set members should add up to 1.

BPAs can be inferred from various evidences using the combination rules of the DS theory. For example, let the polarity score of the word w at time-span T_1, i.e. $Pol(w^1)$, be a positive value $0 \leq s \leq 1$. The BPAs for this evidence can be defined in DS terms as follows:

$$\begin{aligned} &m_{w^1}(\phi) = 0 \\ &m_{w^1}(positive) = s \\ &m_{w^1}(negative) = 0 \\ &m_{w^1}(positive\ or\ negative) = 1 - s \end{aligned} \tag{7}$$

Note that, according to the DS theory, the above evidence only supports the positivity of w and does not convey anything about its negativity. Therefore, the value $1 - m_{w^1}(positive)$ reflects the amount of uncertainty about the status of w at time T_1, i.e. $m_{w^1}(positive\ or\ negative)$. In other words, if the evidence is flawed, w could still be either positive or negative. The uncertainty state of the DS theory is the major characteristic that differentiates this theory from other theories like Bayesian probability theory.

The DS *combination rule* can be defined as follows: Let $m_{w@j}(A)$ be the combined evidence about the polarity of w up to time T_j. The value of $m_{w@j}(A)$ can be computed by combining w's polarity evidences obtained at times $T_1, ..., T_j$, i.e. $\{m_{w^1}(.), ..., m_{w^j}(.)\}$. The DS rule for combining these j evidences is as follows:

$$m_{w@j}(A) = \frac{\sum\limits_{\cap X_i = A} \prod\limits_{1 \leq i \leq j} m_{w^i}(X_i)}{1 - \sum\limits_{\cap X_i = \phi} \prod\limits_{1 \leq i \leq j} m_{w^i}(X_i)} \tag{8}$$

The denominator is the normalization factor that ensures that $m_{w@j}(A)$ is a BPA. We use the above belief values, $m_{w@j}(A), \forall A \in 2^\Theta$, to compute the time accumulated polarity score of w^j up to time-span T_j, $Pol(w@j)$, as follows:

$$\begin{aligned} Pol(w@j) = I * \max[&m_{w@j}(positive), \\ &m_{w@j}(negative), \\ &m_{w@j}(positive\ or\ negative)] \end{aligned} \tag{9}$$

where

$$I = \begin{cases} +1, & \text{if } m_{w@j}(positive) = max \\ -1, & \text{if } m_{w@j}(negative) = max \\ 0, & \text{if } m_{w@j}(positive\ or\ negative) = max \end{cases} \tag{10}$$

The value of $m_{w@j}(positive\ or\ negative)$ indicates the amount of uncertainty that we have about the polarity of w at T_j. Therefore, when this value is maximum, we avoid tagging w as a positive or negative opinion word at T_j and let the future time-spans determine its polarity. We consider any candidate opinion word with a non-zero $Pol(w@j)$ as a new opinion word.

This formulation can tolerate the noise of the polarity scores obtained at different times from the co-occurrence patterns.

3. EXPERIMENTS

We first explain the datasets and some parameter settings and then report the evaluation results.

3.1 Data and Settings

We used three popular opinion lexicons to supply the seeds. In particular, we consider the words that are either labeled as strong in the General Inquirer [8] or subjectivity lexicon [10], or have a positive or negative score of one in SentiWordNet [3] as seeds. Words with the objectivity score of one in SentiWordNet are considered as non-opinion. For SentiWordNet, we only consider the first sense of the words.

We used a large dataset of `Amazon.com` reviews gathered by Jindal and Liu [7] to perform the experiments. This dataset contains more than 5.8M reviews. We only performed the experiments on the reviews from Jan 2000 to May 2006 as there are very few reviews available before 2000 in this dataset. We divided this data into 13 time-spans at 6-monthly time intervals. For sentiment classification of reviews, we balanced the data on the positive and negative reviews.

	T_{10}	T_{11}	T_{12}	T_{13}
top 5, positive				
1	mettle	bros	offerred	healings
2	topnotch	excellance	worshiped	sticklers
3	amassed	muss	soulfully	ubers
4	reigning	earthshattering	excellance	exsperiance
5	fab	soulfully	ubers	dimmu
top 5, negative				
1	irks	gutteral	targetted	plagerized
2	groaner	molested	regretably	dumbledore
3	doomy	derailed	rackets	worsened
4	umph	errie	sqeaky	gimmie
5	maggots	dodged	ozzfest	lamer

Table 1: Top 5 detected words in the latest four time-spans.

In Equation (2), we set the parameter ζ to 3, as suggested by [6], and z to the average sentence length in the above corpus. All the parameters of the optimization framework of Section 2.2 are set to the values of the best performing system as reported in [1].

In all the subsequent experiments, we used the two-tailed paired t-test $p < 0.01$ for significance testing.

3.2 Quality of New Opinion Words

Table 1 shows the top five positive and negative words learned by our method for the latest four time-spans. As it is shown, some misspelled seeds like *excellance*, and *regretably* etc as well as urban words like *fab*, *topnotch*, and *lamer* etc have been accurately detected.

We also evaluated the quality of the discovered new opinion words based on the percentage of such words that are indeed opinion. For this purpose, we manually annotated them as opinion or non-opinion. The average performance was 68.76% (the corresponding Table is removed due to space limit). The annotation shows that our method accurately detected many misspelled seeds as opinion words. In addition, the extracted non-opinion words were mainly the words that frequently co-occurred with one type of seeds (e.g. negative) and rarely co-occurred with the other type. These words were assigned high polarity scores by the optimization framework and consequently labeled as opinion by our system. Such words, though non-opinion, are good polarity indicators. For example, the word "*Dumbledore*" was labeled as negative as it frequently occur in negative reviews (high co-occurrence with negative seeds) but not with positive ones. We noticed that this word refers to a character in the "Harry Potter" series who received many negative opinions against his positive role in the movie. These types of non-opinion words could be used as great polarity indicators.

3.3 Performance of Polarity Inference

For this evaluation, we considered part of the seed dataset as the test data and the rest as training data, and evaluated how the time accumulated polarity improves the polarity of the test seeds computed at different times. We only considered seeds that occur more than 10 times in our review corpus (i.e. 2500+ seeds) and conducted 5-fold cross validation over the seed dataset based on the following evaluation measures:

$$Precision = \frac{N_{correct}}{N_{labeled}},\ Recall = \frac{N_{correct}}{N_{seed}}$$

where $N_{correct}$ is the number of test seeds that were assigned correct polarity (either positive or negative), $N_{labeled}$ is the number of test seeds that were assigned non-zero scores, and N_{seed} is the total number of test seeds. Figure 2 shows the results. At each time T_j, *Spcf* indicates the polarity inference performance of the optimization framework of [1], and *Acm* indicates the performance of the time accumulated polarity computed from the combination of all the polarity evidences obtained up to time T_j, Equation (9). The results show that polarity accumulation through time leads to more accurate polarity inference than the non-time-based method at different times.

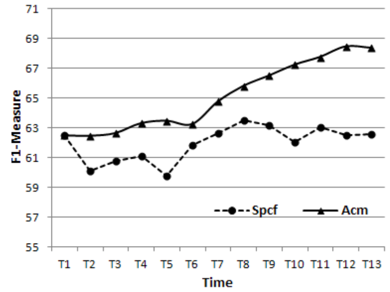

Figure 2: Polarity inference through time.

As Figure 2 shows, the performance of *Acm* increases through time with greater improvements in the latter times. This is because of the availability of more polarity evidences about the test seeds for *Acm* as time passes. However, the performance of *Spcf* depends on the co-occurrence patterns obtained at each time and as Figure 2 shows varies greatly through time. *Acm* significantly outperforms the *Spcf* method by 5.8% on average in F1 score. The difference between the two methods is significant for all T_i, $i \geq 5$.

3.4 Performance of Sentiment Classification

In this section, we study how the learning of new opinion words through time affect the performance of sentiment classification (SC) of reviews. Note that, we use a word-matching-based sentiment classifier instead of popular classifier (like SVM or Naive Bayes) in order to emphasize that the performance improvements come mainly from the quality of the new opinion words. However, we used manually created rules to handle negations as they reverse the sentiment of the opinion words. In particular, we considered words/phrases like *not*, *barely*, *lack of*, and *never* etc, as well as cases that the negation word is not negating the words such as "*not only ... but also ...*", "*last but not least ...*" etc. In total, we compiled 36 negation words and rules.

Given a review, our word-matching-based sentiment classifier computes a sentiment score for the review by summing up the polarity scores of its opinion words. A positive score indicates a positive review, and a negative one indicates a negative review. We expect the performance of this classifier to be better when we use opinion words with higher quality.

Figure 3 shows the performance of SC using seeds and new opinion words. *Seeds* as the baseline indicates the SC performance when we only use seeds to classify reviews of each time-span, while *Seeds+NOW_OPT* and *Seeds+NOW_AC* respectively show the SC performance when we use both seeds and all the new opinion words we learned up to time T_i to classify the reviews of the same time-span T_i. Note that *Seeds+NOW_OPT* uses the most recent polarity score of each new opinion word (obtained from [1]) to perform SC, whereas *Seeds+NOW_AC* utilizes the time accumulated polarity score for this purpose, Equation (9). The results show

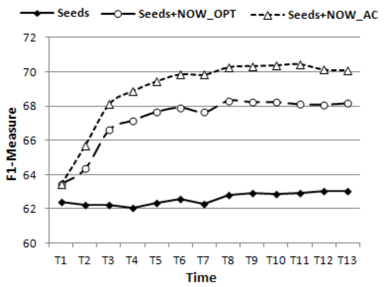

Figure 3: The effect of polarity inference on the performance of sentiment classification.

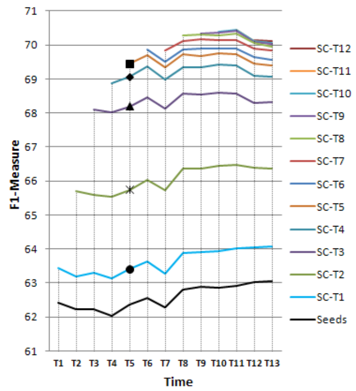

Figure 4: Performance of sentiment classification through time (best seen in color).

that both *Seeds+NOW_OPT* and *Seeds+NOW_AC* significantly outperform *Seeds* for all T_i, $i \geq 2$. This reflects the utility of the new opinion words found for SC. Also, *Seeds+NOW_AC* significantly outperforms *Seeds+NOW_OPT* for all T_i, $i \geq 5$. This shows the effectiveness of the time accumulated polarity scores obtained by DS combination rule.

We also studied the effect of learning more recent new opinion words on the performance of SC. For this purpose, at each time, we used seeds and the current opinion words to perform SC on the current and future reviews. Figure 4 shows the results. Each $SC - Ti$ indicates the performance of SC when we use both seeds and the new opinion words that we learned up to time T_i to perform SC on the current and future reviews, i.e. the reviews of T_k, $\forall k \geq i$. Here, we use the time accumulated polarity scores.

The results show that the SC performance improves as time passes. In other words, each $SC - Ti$ improves the SC performance over the earlier $SC - Tk$, $\forall k \leq i$. For example consider the time T_5. As highlighted in Figure 4, the performance of SC using the new opinion words we learned up to time T_5, i.e. $SC - T5$, is greater than the performance of SC using the new opinion words we learned at earlier times, i.e. $SC - T1$ to $SC - T4$. In other words, the improvement is greater when the sentiment classifier utilizes more recent new opinion words. This is because, in the more recent times, the classifier receives a greater number of new opinion words with more accurate polarity scores due to the existence of more polarity evidences.

4. RELATED WORK

Mining opinion words from user generated content is a crucial prerequisite for effective sentiment analysis. This task includes the detection of new opinion words as well as inferring their polarities. Previous research in this area can be mainly divided into dictionary- and corpus-based approaches. Dictionary-based approaches like [5] utilize dictionaries like WordNet to mine opinion words, whereas corpus-based approaches use synthetic and co-occurrence patterns in text for this purpose [9, 1]. Dictionary-based methods are precise but, in contrast to corpus-based approaches, unable to find informal or so-called urban opinion words.

As a corpus-based approach, Turney and Littman [9] proposed to determine the polarity of a word by comparing its tendency toward positive or negative seeds. Amiri and Chua, (2012) showed that the polarity association among and between seeds and unlabeled words improves the performance of the above method. They showed that both labeled and unhealed data are important for learning the polarity scores.

The difference between this paper and the previous approaches is that we pay attention to ”time” and consider it as another important factor for mining sentiment words.

5. CONCLUSION

We proposed the *interchangeability* method to find new opinion words through time and utilized Dempster-Shafer theory to obtain a time accumulated polarity for each new opinion word. We showed that the time accumulated polarity better reflects the polarity of the opinion words than the polarity obtained at each time. We also showed that mining more recent opinion words leads to better sentiment classification. In the future, we aim to investigate the the effect of the length of the time span on the performance of sentiment terminology mining and sentiment classification.

6. REFERENCES

[1] H. Amiri and T. S. Chua. Mining slang and urban opinion words and phrases from cqa services: An optimization approach. In *Proceedings of WSDM*, 2012.

[2] A. P. Dempster. A generalization of bayesian inference. In *Journal of the Royal Statistical Society*, B(30), 1968.

[3] A. Esuli and F. Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC*, 2006.

[4] S. G. A mathematical theory of evidence. In *Princeton University Press*, 1976.

[5] A. Hassan and D. R. Radev. Identifying text polarity using random walks. In *Proceedings of ACL*, 2010.

[6] A. Islam and D. Inkpen. Semantic text similarity using corpus-based word similarity and string similarity. In *Proceedings of TKDD*, 2(2), 2008.

[7] N. Jindal and B. Liu. Opinion spam and analysis. In *Proceedings of WSDM*, 2008.

[8] P. J. Stone and E. B. Hunt. A computer approach to content analysis: studies using the general inquirer system. In *Proceedings of SPRING*, 1963.

[9] P. D. Turney and M. L. Littman. Measuring praise and criticism: Inference of semantic orientation from association. In *Proceedings of TOIS*, 21(4), 2003.

[10] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of HLT-EMNLP*, 2005.