Research Journal of Applied Sciences Engineering and Technology 4(7): 833-839, 2012
ISSN: 2040-7467

Submitted: December 06, 2011 Accepted: December 26, 2011 Published: April 01, 2012

Identifying Product Features from Customer Reviews using Lexical Concordance

Khairullah Khan and Baharum B. Baharudin
Department of Computer and Information Sciences, Universiti Teknologi PETRONAS Malaysia, City, Tronoh, Perak, Malaysia

Abstract: Automatic extraction of features from unstructured text is one of the challenging problems of Opinion Mining. The trend of getting products and services reputation from online resources such as web blogs and customer feedback is increasing day by day. Therefore efficient system is required to automatically extract products features and the opinion of consumers about all aspects of the products. In this study our focus is on extraction of product features from customer reviews. We have proposed a concordance based technique for automatic extraction of features of product from customer reviews. In our proposed technique we extract patterns of lexical terms using concordance for candidate features extraction and identify features by grouping. The proposed grouping algorithm is used to remove irrelevant features. We conducted experiments on different products reviews and compared our results with existing methods. From empirical results we proved the validity of the proposed method.

Key words: Concordance, feature extraction, feature grouping, opinion mining

INTRODUCTION

Opinion has great importance in decision making. Opinion play very significant role in our day to day life. It helps in suiting, molding and fashioning our decisions, when we observe and analyze. Now the world has become a global village due to the internet. Through internet we can share our knowledge and opinion for making right decision well in time. The problem is how to extract the reviewer concerns from unstructured review text. In order to solve this problem text mining techniques plays a vital role. OM is the computational study of opinions, sentiments and emotions expressed in text (Bo Pang, 2008).OM is carried out using Natural Language Processing (NLP) and statistical comptational techniques to mine opinion(Khan *et al*., 2009). OM differs from typical text mining as it is concerned with the analysis of opinion expressed in text instead of typical text analysis (Bo Pang, 2008; Liu *et al*., 2005; Olivas, 2009). Opinions can be gathered in two different formats i.e structured and unstructured Bo Pang, 2008).The structured opinions are gathered through questioners/feedback form which bound the user to give answer of the given questions in yes or no and are stored in tabular form. Structured opinions can be easily processed and analyzed. Unstructured opinion is collected in the form of plain text for example comment box, blogs and etc. Information retrieval from plain text is more complex as compared to the structured due to high frequency of irrelevant terms, sparseness and ambiguity. Thus unstructured opinion mining is more complex and involves in-depth processing. For OM different components are needed to extract. The following are the main components of Opinion (Kobayashi, 2007; Liu *et al*., 2005; Su, 2007).

- Target object about which opinion is expressed.
- **Features:** The characteristics and aspects of the object discussed by the reviewer in the reviews.
- **Opinion:** Expression of opinion holder about the features of the products.
- **Opinion holder:** The review who has given the opinion.

One of the most important and complex problem is identification of features of the target object about which opinion is expressed. In this study our focus is identification and extraction of features of concern from reviews from unstructured text. The extraction of individual features of object is important to know about every aspects of the product. Based on our literature study we found that processing and summarizing opinions about individual features plays a key role. For example we have a set of product features, in order to show the customer opinion about individual features it can be presented in a summarized way as shown in Fig. 1. From this visualization one can easily compare the individual features of the product. The in-depth analysis of every aspect of a product based on consumer opinion is important for both consumers and manufacturer. In order to compare the reviews it is required to automatically

Corresponding Author: Khairullah Khan, Department of Computer and Information Sciences, Universiti Teknologi PETRONAS Malaysia, City, Tronoh, Perak, Malaysia

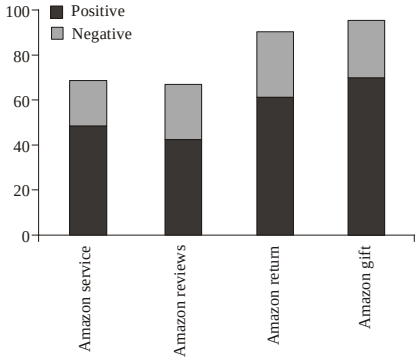

Fig. 1: Example of opinions summarization of individual features of target object

identify and extract those features which are discussed in the reviews. Thus features mining of products are important for opinion mining and summarization. The task of features mining provides a base for opinion summarization (Feldman *et al*., 2007). A consumer can have different opinion about different aspects/features of a product. One can express a positive opinion about one feature and a negative about another feature of single product. Thus it is not necessary that every aspect is liked/ or disliked by everyone.

Our main contribution in this study is the grouping algorithm based on the lexical patterns produced. The aim of this algorithm is to remove irrelevant features and get list of concerned features of the product.

LITERATURE REVIEW

As mentioned in this study we concentrate on feature extraction of products. In this section we have presented a background study of existing work in order to review existing techniques and highlight the scope and motivation of our proposed study. Several approaches are available for this sub task of OM. These approaches can be broadly divided into two major categories: supervised and unsupervised. The supervised learning approaches are based on manually labeled text. In this approach a machine learning model is trained on manually labeled data to classify and predict features in the reviews. Although supervised techniques provide good results in feature extraction, it requires a manually and skill-oriented process which is time consuming and laborious.

Also, the effectiveness of supervised model depends on the training set whereas preparation of training sets requires human expertise. Generally the most widely used supervised techniques are Decision Tree, K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Neural Network, and Naïve Bayesian Classifier (Weiss *et al*., 2005). The application of supervised learning method has been reported in different papers for features

Table. 1: Summary of consumer review data set

Pattern	RE
JJNN	\w*/JJ\w* \w*/NN\w*\W+
JJNNNN	\w*/JJ\w* \w*/NN\w* \w*/NN\w*\W+
NN	\w*/NN\w*\W+
NNNN	\w*/NN\w* \w*/NN\w*\W+
NNNNNN	\w*/NN\w* \w*/NN\w* \w*/NN\w*\W+

extraction from plan text (Abbasi *et al*., 2008; Pang *et al*., 2002; Simmons, 2004)

On the other hand unsupervised techniques do not require labeled data and automatically predict product features based syntactic rules and patterns in textual data. The most popular and widely used unsupervised techniques are Clustering, Singular Value Decomposition (SVD), and Component Analysis (Kamber, 2000; Weiss *et al*., 2005).Unsupervised learning for product features extraction has been reported in different papers (Chen *et al*., 2010a, 2010b; Wong and Lam, 2009). For example (Yi *et al*., 2003), have used different patterns of base noun phrases for candidate feature extractions and then applied likelihood ratio test to extract relevant features from candidate list. For candidate features extraction we are also using noun phrases as shown in Table 1 and likelihood ratio for relevance scoring. But in our approach we further improved the results using features refinement and grouping. (Liu *et al*., 2008) in their work have proposed association rule mining technique Agrawal and Srikant, 1994)based on noun phrases. The same technique is applied by (Wei, 2010) with enhancement of semantic based refinement of extracted features. The Association Rule Mining depends on frequency of associated patterns of terms and do not consider relevancy to the topic concerned therefore its performance degrades when the documents have irrelevant frequent terms. (Ferreira *et al*., 2008) have compared the association rule mining algorithm with likelihood ratio test for product features extraction and reported that likelihood performs comparatively well, when the reviews have irrelevant frequent features. Normally in reviews the reviews are not bound and can write about different objects in a same review. Therefore it is needed to relate the features to the target object or domain. The same problem can occur with the likelihood ratio test (Dunning, 1993) because it totally depends on documents to distinguish their features. Thus it is necessary to remove irrelevant features by grouping them and to find relatedness of each group of features to the domain of the target object. In our proposed framework we are grouping and binding the features to the target domain which is being represented in the documents. In order to group the features we have proposed an algorithm based on the co-occurrence of terms in the left and right context. The groups are bind to the target domain using is-a relationship. Thus our method first checks the relevancy in the documents and then relates the features groups to the domain of the target

object. In this way we get a refined collection of features related to the target object.

We have proposed a framework for feature identification and extraction which consist of few steps as given in our proposed architecture and explained below. To the best of our knowledge no such framework has been reported which exactly match our proposed framework. Based on empirical evaluation we found that our proposed framework performs comparatively well.

PROPOSED FRAMEWORK

In this study we have proposed a framework for extraction of product features from plain text. Our proposed technique is basically depends on noun phrase concordance and relevance scoring therefore we call it Concordance based Feature Extraction (CFE). This technique is using regular expressions to extract candidate features through lexical patterns in left and right context of a term; and find relevancy of the candidate features to the target object using LR technique. Our proposed framework is starting from preprocessing of consumer reviews, where the preprocessing task includes part-of-speech tagging (POS), and stemming. Then we use relevance scoring (Dunning, 1993) to identify relevant features. The grouping step of proposed framework is used for find exact features of target object. Figure 2 shows the overall process of our proposed framework. In the following subsections, we explain the detailed implementation of each step.

Preprocessing: The CFE technique begins with preprocessing of input consumer reviews, which consists of POS tagging and stemming. In this we have used widely reported software from existing literature the "Stanford Tagger" (Manning, 2003) for POS and Porter Stemmer (Porter, 1980) for Stemming.

Candidate feature extraction: After preprocessing we extract noun phrases as candidate features through R.E as given in Table 1.

Features refinement: Before the extraction of features we use the following two refinement steps to remove irrelevant terms from candidate features.

Refining through noun modifier: Along stop words we use some nouns which are used to modify another noun but do not represent any specific feature. For example nothing/NN, anything/NN, excuse/NN etc although categorized as noun but do not represent features. They are required to be removed from the candidate list to avoid irrelevant features. We collected a seed list of such noun modifier with the consultation of English language experts and use thesaurus dictionary to expand the list from its synonyms and use in this refinement step.

Table. 2: Summary of consumer review data set

Product	N.o of manually tagged product features	N.o of review sentences
Apex	110	739
Cannon	100	597
Creative	180	1716
Nikon	74	346
Nokia	109	546

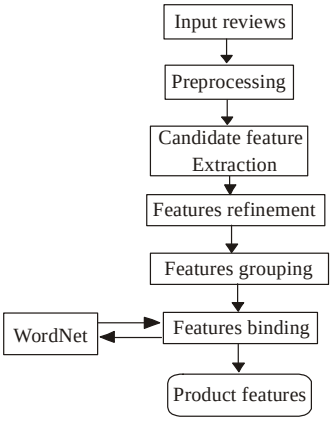

Fig. 2: Overall process of concordance based product feature extraction (CFE) technique

Refining through evaluative adjectives: As mentioned in Table 2 there are two types of noun phrases i.e the phrases which contain only nouns and phrases which contain nouns and adjectives. In this step we process only those candidate features which has adjective with nouns. e.g in the phrase "Excellent/JJ Fan/NN", only the fan can be a feature, however in a phrase "digital/JJ camera/NN" both terms represent a feature. In the first phrase the excellent is an evaluative adjective while the fan is actually a feature of the product. On the other hand the phrase "digital/JJ camera/NN" is a feature but the digital is not evaluative adjective. In order to refine this type of candidate features we use seed list of positive and negative terms collected (Hu and Liu, 2004).

Features grouping: The last step of our framework for feature extraction is to group the related features. We further refine the obtained list of features by grouping them. Our proposed algorithm is based on context of consecutive terms. The main advantage of features grouping is to concise features and to differentiate different aspect of the object of concern, and to further refine the features by removing irrelevant group of features. The exact grouping of features is still challenging problem. For example we have the following review sentences:

- Yesterday, I bought a Nokia phone and my girlfriend bought a moto phone
- We called each other when we got home
- The voice on my phone was not clear

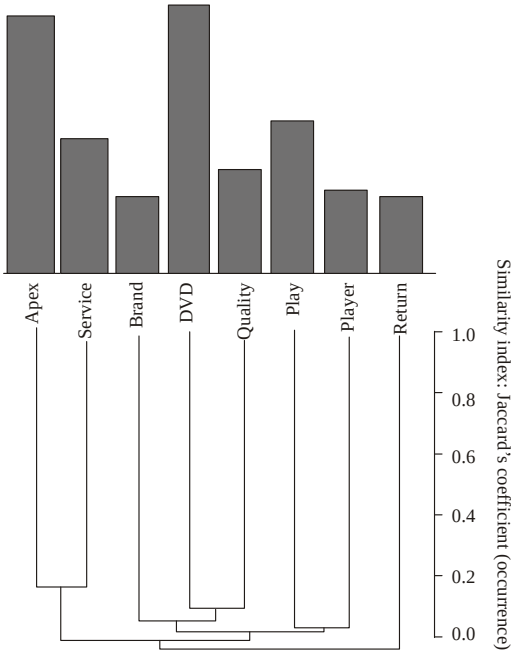

Fig. 3: Grouping feature by concordance

- The camera was good
- My girlfriend said the sound of her phone was clear
- I wanted a phone with good voice quality
- So I was satisfied and returned the phone to BestBuy yesterday

In these sentences features of different categories can be found. For example the features can be grouped as:

- Group of features related to phone
- Group of features related to camera
- Group of features related to company
- Group of features related to seller and buyer

We proposed the following two steps process for feature grouping.

Grouping features: In order to group N-Gram features we have proposed a novel algorithm which is based on concordance of the unigram features with multi features. We consider left and right context for extraction of multi words features as already given in section B. Comparatively N-gram feature provides better concept as compare to Uni-Gram. In this step we have proposed a novel algorithm based on context of consecutive N-Gram words belongs to extracted features. The sample Dendrogram of grouping is given in the Fig. 3.

Definition: Let we have a set of features F extracted from documented D, M is subset of F having Multi Words Features, then for every f ϵ F, we find the group of feature f using the model given in Eq. (1).

$$G(f)=\begin{cases} g, & g \in g \cap M \\ gn & otherwise \end{cases} \quad (1)$$

where gn represent a new group. The process is iterated for all features in F as shown in the following algorithm.

Algorithm for features grouping: Let Fg represents feature group, F represents Feature sets extracted through likelihood ratio, Mg represents multi features group, and Tg represents temporary group. Then the grouping is done using the following algorithm.

1. <Initialize Fg = Ø>
2. <For each Fi in F>
3. <Initialize Mg = Ø>
4. <Initialize Tg = Ø>
5. <for each Mi in M-Fi>
6. <If Fi exist in Mi then>
7. <Mg = Mg U Fi>
8. Else
9. <Tg = Tg U Fi>
10. </IF>
11. <Fg = Fg U Mg>
12. </For>
13. <For Each tg in Tg>
14. <For each fg in Fg>
15. <If tg exist in fg then>
16. <Fg = Fg U ng>
17. <Ng = Ng-ng>
18. </IF>
19. </For>
20. </For>

Grouping uni-gram features: Some group of features will not have multi word features. For such type of Uni-Gram features we propose dictionary based grouping based on semantic relatedness. This type of grouping is reported by Hu and Liu in Ref. (Hu and Liu, 2004) for grouping opinionated features.

Features binding to the target object: In this step we propose the identification of target object based on is-a relationship through WorldNet (Riesenfeld, 1998) dictionary. For finding is-a relationship we use the following algorithm.

- For each group G in groups
- For each feature F in G
- IF F exists in Hyponym of G then
- Return F

Empirical evaluation: We have performed empirical evaluation and compared our results with two existing methods based on association rule mining for extraction

Table 3: Summary of extracted features by proposed CFE

Product	Candidate features	Total extracted	Correctly extracted
Apex	203	111	80
Cannon	198	99	75
Creative	320	180	117
Nikon	180	72	55
Nokia	205	107	79

Table 4: Precision and recall of proposed CFE

Product	MA	TE	CE	P%	R%	F %
Apex	110	111	80	54	72	61.71
Cannon	100	99	75	49	76	59.58
Creative	180	180	117	44	65	52.48
Nikon	74	72	55	48	76	58.84
Nokia	109	107	79	58	73	64.64

of product features. These techniques have been reported in the work of (Hu and Liu, 2004) and then by (Wei, 2010) with improvement using semantic based features refinement. In our work we have used Likelihood Ratio test and lexical Concordance. In our framework we have proposed features grouping using multi word phrase to remove irrelevant features. In the following subsections the detail of our experiments including the data collection and evaluation criteria is given. Subsequently, we have discussed some important evaluation results.

Data collection: We use the consumer review data set of five products which was first used by(Hu and Liu, 2004) and is available online from home page of the author. This data contains reviews of five different products: Apex AD2600 Progressive-scan DVD player, Canon G3, Creative Labs Nomad Jukebox Zen Xtra 40 GB, Nikon Coolpix 4300, and Nokia 6610. The detail of the data is given in Table 1.

Evaluation criteria: To evaluate the effectiveness of CFE, we compare our results with the bench marked results by (Wei, 2010). They have compared their results with manually tagged product features extracted by (Hu and Liu, 2004) . In their work they have adopted fuzzy matching method i.e. if one or more words in an extracted product feature are exactly the same as some word(s) in a manually tagged product feature, then they consider it as a product feature. For example, ''DVD player'' was identified as a product feature in some review sentences, and ''player'' was tagged as a product feature in others. The count of extracted features by our proposed technique is given in Table 3.

We use precision, recall and F-score to measure the effectiveness of a product feature extraction technique. For a product i, its precision, recall and F-Score are calculated using the following Eq. (2), (3) and (4), respectively.

$$P_i = \frac{c_i}{e_i} \quad (2)$$

Table 5: Experimental results of comparative study

	ARM			SFE			CFE		
Product	P	R	F	P	R	F	P	R	F
Apex (%)	51	60	55.14	52	70	59.93	54	72	61.91
Cannon (%)	51	63	56.43	49	75	59.05	49	76	59.68
Creative (%)	37	56	44.59	44	65	52.48	44	65	52.69
Nikon (%)	51	68	58.14	47	76	58.30	48	76	58.76
Nokia (%)	50	58	53.33	57	73	63.51	58	73	64.97

Table. 6: Example of features refinement by grouping through concordance of lexical categories

Doc	Phrase
Apex	Rebat
Apex	Courtesy
Apex	Crap
Apex	Lips

$$R_i = \frac{c_i}{m_i} \quad (3)$$

$$F_i = 2 * \frac{P_i * R_i}{P_i + R_i} \quad (4)$$

where P_i, R_i, And F_i denote Precision, Recall, and F-Score respectively. And moreover c_i is the number of manually tagged product features of product i that are correctly extracted by the technique under examination, e_i is the number of product features of product i extracted by the technique, and m_i is the number of manually tagged product features of product i. The precision, recall and F-Scores of our proposed technique are given in Table 4. Here MA represents manually annotated features, TE represents total extracted features by proposed technique, CE represents correctly extracted features, P represents precision, R represents Recall, and F represents F-Scores.

Comparative study: Based on experimental evaluation we found that average F-scores of our proposed CFE techniques is higher than the existing approaches. As shown in Table 5 the average F-score of Association Rule Mining approach (ARM) is 53.53, while Semantic Based Feature Extraction (SFE) provides 58.65 and our proposed Concordance based Feature Extraction (CFE) provides 59.60 which are higher than existing techniques.

Effect of grouping on precision and recall: We compared our results before and after the grouping algorithm and found that our proposed grouping algorithm significantly improved the results. About 15 to 18% improvement was achieved in precision however the recall becomes slightly low. For example the group of irrelevant features was identified in groups of DVD player which were not detected by likelihood ratio test are shown in Table 6.

CONCLUSION AND FUTURE WORK

In this study we have proposed a unified framework to identify and extract product features from unstructured reviews. Our system can assist individual customer and consumers, vendors, and production companies in decision making. The problem that we have addressed is

the extraction of relevant features and to group those features in relevant categories, because a review can have irrelevant features that can cause deficiency of the system. Our proposed framework has few steps in which we combined linguistic techniques with our novel algorithm for feature identification. For grouping features we have proposed our algorithm which is based on co-occurrence of features using left and right context. We conducted extensive experiments on datasets of different products to demonstrate the performance of our proposed framework and to compare our results with the existing techniques. Based on experimental results we found that our proposed framework provides good results. We intend to extend our framework for extraction of evaluative expression and other opinion components from unstructured text. One possible extension is to incorporate existing resources of lexicon for extraction of evaluative expressions and relationship between features and evaluative terms. The framework can further be improved by prior domain knowledge.

REFERENCES

Abbasi, A., H. Chen and A. Salem, 2008. Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums. ACM Trans. Inf. Syst., 26(3): 1-34.

Agrawal, R. and R. Srikant, 1994. Fast Algorithms for Mining Association Rules in Large Databases. Paper presented at the Proceedings of the 20th International Conference on Very Large Data Bases.

Bo Pang, L.L., 2008. Opinion mining and sentiment analysis. Foundations Trends Inf. Retrieval, 2(1-2): 135.

Chen, H.W., K.R. Lee, H.H. Huang and Y.H. Kuo, 2010a. Unsupervised subjectivity-lexicon generation based on vector space model for multi-dimensional opinion analysis in blogosphere. Paper presented at the Proceedings of the 6th International Conference on Advanced Intelligent Computing Theories and Applications: Intelligent Computing.

Chen, H.W., K.R. Lee, H.H. Huang and Y.H. Kuo, 2010b. Unsupervised Subjectivity-Lexicon Generation Based on Vector Space Model for Multi-Dimensional Opinion Analysis in Blogosphere Advanced Intelligent Computing Theories and Applications. In: Huang, D.S. Z. Zhao, V. Bevilacqua and J. Figueroa, (Eds.), Springer Berlin, Heidelberg, 6215: 372-379.

Dunning, T., 1993. Accurate methods for the statistics of surprise and coincidence. Comput. Linguist., 19(1): 61-74.

Feldman, R., M. Fresco, J. Goldenberg, O. Netzer and L. Ungar, 2007.extracting product comparisons from discussion boards. Paper presented at the Proceedings of the 2007 Seventh IEEE International Conference on Data Mining.

Ferreira, L., N. Jakob and I. Gurevych, 2008. A comparative study of feature extraction algorithms in customer reviews. Paper presented at the Semantic Computing, 2008 IEEE International Conference on.

Hu, M. and B. Liu, 2004. Mining and summarizing customer reviews. Paper presented at the Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining.

Kamber, J.H.A.M., 2000. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Khan, K., B.B. Baharudin, A. Khan and F. E-Malik, 2009. Mining opinion from text documents: A Survey. Paper presented at the Digital Ecosystems and Technologies. DEST '09. 3rd IEEE International Conference on.

Kobayashi, 2007. Opinion Mining from Web documents: Extraction and Structurization. Nara Institute of Science and Technology.

Liu, B., M. Hu and J. Cheng, 2005. Opinion observer: Analyzing and comparing opinions on the Web. Paper presented at the Proceedings of the 14th international conference on World Wide Web.

Liu, H., H. Yang, W. Li, W. Wei, J. He and X. Du, 2008. A system for online review structurization. Paper presented at the Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining.

Manning., D.K.A.C.D., 2003. Accurate Unlexicalized Parsing. Paper presented at the Proceedings of the 41st Meeting of the Association for Computational Linguistics.

Olivas, E.S., 2009. Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods and Techniques.

Pang, B., L. Lee and S. Vaithyanathan, 2002. Thumbs up?: Sentiment classification using machine learning techniques. Paper presented at the Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, Vol. 10.

Porter., M.F. 1980, An algorithm for suffix stripping. Program, 14(3): 130–137.

Riesenfeld, M.M.S.A.R.F., 1998. WordNet: An Electronic Lexical Database. Paper presented at the 11th Eurographics Workshop on Rendering.

Sparck Jones, Karen, and Peter Willet, 1997, Readings in Information Retrieval, San Francisco: Morgan Kaufmann, ISBN: 1-55860-454-4.

Simmons, R.G.A.H., 2004. Predicting the end-price of online auctions. Paper presented at the International Workshop on Data Mining and Adaptive Modelling Methods for Economics and Management.

Su, Q., 2007. Mining feature-based opinion expression by mutual information approach. Inter. J. Comput. Proc. Oriental Lang., 20(2-3): 137-150.

Wei, 2010. Understanding what concerns consumers: A semantic approach to product features extraction from consumer reviews. Info Syst E-Bus Manage., 8: 149-167.

Weiss, S.M., N. Indurkhya, T. Zhang and F. Damerau, 2005. Text Mining. In 2005 (Eds.), Predictive Methods for Analyzing Unstructured Information.

Wong, T.L. and W. Lam, 2009. An unsupervised method for joint information extraction and feature mining across different Web sites. Data Knowledge Engineering, 68(1): 107-125.

Yi, J., T. Nasukawa, R. Bunescu and W. Niblack, 2003. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. Paper presented at the Data Mining, 2003. ICDM 2003. Third IEEE International Conference on.