Preprint of: Na, J.C., Sui, H., Khoo, C., Chan, S., & Zhou, Y. (2004). Effectiveness of simple linguistic processing in automatic sentiment classification of product reviews. In I.C. McIlwaine (Ed.), *Knowledge Organization and the Global Information Society: Proceedings of the Eighth International ISKO Conference* (pp. 49-54). Wurzburg, Germany: Ergon Verlag.

Effectiveness of Simple Linguistic Processing in Automatic Sentiment Classification of Product Reviews

Authors:
Jin-Cheon Na[1] (tjcna@ntu.edu.sg)
Haiyang Sui[1] (PG01712719@ntu.edu.sg)
Christopher Khoo[1] (assgkhoo@ntu.edu.sg)
Syin Chan[2] (asschan@ntu.edu.sg)
Yunyun Zhou[1] (ZHOU0015@ntu.edu.sg)

1. Authors' address:
Division of Information Studies
School of Communication & Information
Nanyang Technological University
31 Nanyang Link, Singapore 637718
Tel: (65) 6790-5011 Fax: (65) 6792-7526

2. Authors' address:
Division of Computer Communications
School of Computer Engineering
Nanyang Technological University
31 Nanyang Link, Singapore 637718
Tel: (65) 6790-5748 Fax: (65) 6792-7526

Jin-Cheon Na, Haiyang Sui, Christopher Khoo, Syin Chan, Yunyun Zhou
School of Communication & Information
Nanyang Technological University, Singapore

Effectiveness of Simple Linguistic Processing in Automatic Sentiment Classification of Product Reviews

Abstract: This paper reports a study in automatic sentiment classification, i.e., automatically classifying documents as expressing positive or negative sentiments/opinions. The study investigates the effectiveness of using SVM (Support Vector Machine) on various text features to classify product reviews into *recommended (positive sentiment)* and *not recommended (negative sentiment)*. Compared with traditional topical classification, it was hypothesized that syntactic and semantic processing of text would be more important for sentiment classification. In the first part of this study, several different approaches, *unigrams (individual words)*, *selected words (such as verb, adjective, and adverb)*, and *words labeled with part-of-speech tags* were investigated. A sample of 1,800 various product reviews was retrieved from Review Centre (www.reviewcentre.com) for the study. 1,200 reviews were used for training, and 600 for testing. Using SVM, the baseline unigram approach obtained an accuracy rate of around 76%. The use of selected words obtained a marginally better result of 77.33%. Error analysis suggests various approaches for improving classification accuracy: use of *negation phrase*, making inference from superficial words, and solving the problem of *comments on parts*. The second part of the study that is in progress investigates the use of *negation phrase* through simple linguistic processing to improve classification accuracy. This approach increased the accuracy rate up to 79.33%.

1. Introduction

Research in *Automatic Text Classification* seeks to develop models for assigning category labels to new documents or document segments based on a training set of documents that have been pre-classified by domain experts. Most studies of automatic text classification have focused on "topical classification", i.e., classifying documents according to various subjects (e.g., education vs. entertainment). This study is in the area of "Sentiment Classification" – automatically classifying documents according to the overall sentiment expressed in them. In particular, this study investigated the application of machine-learning methods for classifying product reviews into two categories: *recommended (positive sentiment)* and *not recommended (negative sentiment)*.

Automatic sentiment classification is useful in many areas. It can be used to classify product reviews into positive and negative, so that potential customers can have an overall idea of how a product is perceived by other users (Turney, 2002; Pang, Lee & Vaithyanathan, 2002; Dave, Lawrence & Pennock, 2003). It can also be used to classify Web articles into positive or negative comments, enabling users to browse Web pages more efficiently. Moreover, the technique can be used for filtering out email messages with impolite or abusive words (Spertus, 1997). In the area of social science research, it can be used to categorize news articles into positive and negative views, according to various research purposes (Semetko & Valkenburg, 2000; Lind & Salo, 2002).

Though machine-learning techniques have long been used in topical text classification with good results, they are less effective when applied to sentiment classification (Pang, Lee & Vaithyanathan, 2002). Sentiment classification is a more difficult task compared to traditional

topical classification, which classifies articles by comparing individual words (unigrams) in various subject areas. In sentiment classification, unigrams may not be enough for accurate classification. For instance, the two phrases "you will be disappointed" and "it is not satisfactory" do not share the same words, but both express negative sentiments.

In the first part of the study, we investigated several different approaches to perform the classification using different text features: *unigrams (individual words), selected words (such as verb, adjective, and adverb)*, and *words labeled with part-of-speech tags*. Then we analyzed the product reviews that were wrongly classified by the SVM model to identify the sources of error and directions for improving the automatic classification. The second part of the study that is in progress investigates the use of *negation phrase* through simple linguistic processing to improve classification accuracy.

2. Research Method

2.1 Sampling

Using a program, product reviews were automatically downloaded from Review Centre (www.reviewcentre.com, 2003), which hosts millions of product reviews by consumers. After filtering out blank Web pages, a sample of 1,800 product reviews was systematically selected, comprising 900 positive reviews and 900 negative reviews. The sample was divided into a training set of 1,200 reviews (600 positive and 600 negative) for developing the classification model, and a test set of 600 reviews (300 positive and 300 negative) for evaluating the accuracy of the model. The majority of reviews are of mobile phones and electronic equipments.

Review Centre rates product reviews using a 10-star rating system. In this study, reviews with 7 stars or above are coded as *recommended (positive)*, while reviews with 4 stars or below are *non-recommended (negative)*. This assumption was generally correct, but there was some inconsistency between the ratings and reviewers' comments (see the Error Analysis section). The aim of the classification model is to predict from the natural language text of the review whether the review is coded as recommended or non-recommended.

2.2 Pre-Processing

The texts of the reviews were tokenized and the words extracted were stemmed using the Conexor parser (Tapanainen & Järvinen, 1997). Each review was converted into a vector of terms (i.e. words) with term weights, indicating the importance of each term in the review. Three weighting schemes were investigated: *Term Presence (binary weighting), TF (Term Frequency)*, and *TFIDF (Term Frequency Inverse Document Frequency)*.

Term Presence (binary weighting) has the value 1 if the term exists in the review, 0 otherwise. *Term Frequency (TF)* uses the frequency of the term in the review as the weight. The *TFIDF* weight has been used in many studies on topical text classification, and is defined by the formula: $TF \times \text{Log}(\frac{N}{DF})$

where *TF* is the number of times the term occurs in the current review document, *N* the number of reviews in the training set, and *DF* the document frequency – the number of reviews in the training set containing the term.

2.3 Machine-learning Methods

A machine-learning method, Support Vector Machine (SVM), was used in this study. SVM has been applied to text classification in Joachims's study (1998), and later used in many

studies (Joachims, 1999; Schohn & Cohn, 2000). The core idea is to find a hyperspace surface H, which separates positive and negative examples with the maximum distance. Yang (1999) claimed that SVM and k-NN methods were significantly better than other classifiers. Sebastiani (2002) reported that SVM delivered very good performance in some experiments. In our study, SVM^{light} (www.svmlight.joachims.org, 2003), a publicly available SVM program, was used for automatic review classification.

In a previous study, some of the authors of this paper applied Decision Tree induction (Quinlan, 1983) to identify useful words for sentiment classification (Sui, Khoo, & Chan, 2003). Many studies, including our previous work, have found SVM to perform better than Decision Tree on text classification (Joachims, 1998; Yang & Liu, 1999). However, the result model of decision tree is easy to interpret and can be converted to IF-THEN rules.

2.4 Approaches Investigated

Different kinds of linguistic features were investigated in developing the classification models:

- **Baseline (Unigram)** -- simply used all the individual stemmed words (unigrams) that appeared in product reviews.
- **Selected words (such as verb, adjective, and adverb) --** Conexor parser was used to tag individual words with part-of-speech tags. Only words with specific part-of-speech, such as verb, adjective, and adverb, were used in developing the classifier.
- **Words labeled with part-of-speech (POS) tags** -- the individual words were combined with their POS tags. For instance, the words "better:adjective" and "better:verb" were considered different terms.

3. Results

Table 1 lists the results of the various approaches attempted in this study.

ID	Approach	Selected Terms	Term Weighting	DF	Terms labeled with POS tags	Negation	Accuracy
1	Unigram with TF	All	TF	3	No	No	74.17%
2	Unigram with Presence	All	Presence	3	No	No	75.50%
3	Unigram with TFIDF	All	TFIDF	3	No	No	76.50%
4	Unigram with TFIDF and DF = 1	All	TFIDF	1	No	No	74.17%
5	Unigram labeled with POS	All	TFIDF	3	Yes	No	75.83%
6	Unigram with selected words (V, A, Adverb)	Verb, Adjective, Adverb	TFIDF	3	No	No	77.33%
7	Unigram with selected words (N, V, A, Adverb)	Noun, Verb, Adjective, Adverb	TFIDF	3	No	No	75.50%

Table 1. Various approaches and results

The use of unigrams, the simplest approach, obtained 75.39% accuracy (average of ID1, ID2, and ID3). The *TFIDF* weighting that is effective in traditional topical text classification performed a little better than *TF* and *Presence* when applied to sentiment classification in this study (ID3). DF (Document Frequency) was used to retain (or consider) only words that occur in at least the specified number of documents in the training set. In general, the value 3 for DF performed a little better than the value 1 (ID3 and ID4).

Limiting the terms to just verbs, adjectives and adverbs improved the accuracy rate: 77.33% (ID6). This supports the hypothesis that positive and negative sentiments are expressed mostly through verbs, adjectives and adverbs. But when we included nouns in addition to verbs, adjectives, and adverbs, the accuracy rate degraded a little bit (ID7). The use of additional part-of-speech information did not improve results, possibly because it increased the number of dimensions (each word is subdivided into different part-of-speech) and reduced the term weight (such as TF) for each term (ID5).

4. Error Analysis

Generally, when applied to topical text classification, the accuracy of SVM is above 85% (Joachims, 1998; Yang & Liu, 1999). Thus we reapplied the training set to the learned SVM model (used *unigram, DF = 3,* and *TF* options) to identify the sources of error and directions for improving the automatic classification. Out of 1,200 training set documents, the total number of wrongly classified documents was 87.

Reasons	Number of Documents
Negation phrase	34
Comments on parts	25
Need inferencing	17
Inconsistency between rating and comments	13
Comments on other products	10

Table 2. Error analysis

The possible reasons for failure in automatic classification are summarized in Table 2 and explained as follows (note that some documents are counted multiple times since they have more than one reason for misclassification):

1. **Negation phrase.** Negation phrases in the reviews affected the effectiveness of the simple unigram-approach classifier. For instance, the sentence *"I'd never regretted purchasing it"* is actually a positive comment. However, the unigram approach treats *"never"* and *"regretted"* as separate negative words. This seems to be one of the most common problems in sentiment classification.
2. **Comments on parts.** Sometimes, though the reviewer comments negatively on parts of the product, he is actually satisfied with the product as a whole, e.g., *"The best phone I've had yet. The ONLY bad point is that …"* .
3. **Need inferencing.** Some comments are complex and need inferencing to identify the sentiment classification. The sentence *"if the price dropped, the company would be surprised how it would sell"* contains no apparent positive or negative words.
4. **Inconsistency between rating and comments.** In 13 cases, there is no obvious relation between the reviewer's comments and the number of stars given. For instance, the comment *"Good if you constantly listen to music on the move. This phone is still the best looking phone on the market"* is apparently positive; however the reviewer gave it 3 stars.

5. **Comments on other products.** The reviewer uses indicative words to comment on or make comparisons with other related products. For example, *"8210 is better. More valuable"*.

In addition, some reviews are too short to be classified accurately (23 documents are no longer than two lines). For example, the comment *"This is an OK phone but slow"* is difficult to classify without more context.

As shown in Table 2, the errors attributed to *Negation phrase*, *Comments on parts,* and *Need inferencing* account for a large portion of the wrong classifications. We are currently investigating the use of simple linguistic processing to address the problems of *negation phrase*. Each negation and its adjacent words are combined to generate a new composite term (i.e., negation phrase). To extract negation phrases, we use syntactic patterns, such as *"<Verb> - <Negative Particle> - <Verb>"* and *"<Verb> - <Negative Particle> - <Adverb> - <Adjective>"*. Table 3 lists some samples of *negation phrases* extracted automatically. For instance, "Do not buy" occurs in 34 reviews out of 1,200 training reviews.

Negation Phrases	DF	Negation Phrases	DF
Do not buy	34	Will not regret	3
Do not work	24	Be not as good as	3
Would not recommend	14	Would not buy	2
Do not want	14	Be not very impressed	2
Do not like	9	Be not happy	2
Be not worth	6	Not so bad	2
Not bad	5	Do not purchase	1
Not the good	5	Do not dislike	1
Have not regret	4	Not so good	1
Will not work	4	Not too bad	1
Do not recommend	3	Be not a good choice	1

Table 3. Negation phrases

Table 4 lists the results of the *negation phrase* approaches where *negation phrases* were treated as unique terms. This approach improved the accuracy rate up to 79.33%.

ID	Approach	Selected Terms	Term Weighting	DF	Terms labeled with POS tags	Negation	Accuracy
1	Unigram with negation phrase and DF = 3	All	TFIDF	3	No	Yes	78.33%
2	Unigram with negation phrase and DF = 1	All	TFIDF	1	No	Yes	79.33%

Table 4. Negation phrase approaches and results

5. Conclusion

The use of "*negation phrase"* through simple linguistic processing improved classification accuracy. This suggests that the simple unigram approach for sentiment classification is not good enough. We plan to explore further syntactic and semantic processing and inferencing in order to improve the accuracy rate.

References

Dave, D., Lawrence, S., and Pennock, D. M. (2003). Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th International World Wide Web Conference,* Budapest, Hungary, May 20-24, 2003, 519-528.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of 10th European Conference on Machine-learning,* Chemnitz, Germany, April 21-24, 1998, 137-142.

Joachims, T. (1999). Transductive inference for text classification using support vector machines. In *Proceedings of 16th International Conference on Machine-learning*, Bled, Slovenia, June 27-30, 1999, 200-209.

Lind, R. A. and Salo, C. (2002). The framing of feminists and feminism in news and public affairs programs in U.S. Electronic Media. *Journal of Communication*, *52*(1), 211-228.

Pang, B., Lee, L., and Vaithyanathan, S. (2002) Thumbs up? Sentiment classification using machine-learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, Philadelphia, PA, U.S.A., July 6-7, 2002, 79-86.

Quinlan, R. (1983). Learning efficient classification procedures and their application to chess end games. *Machine Learning: An Artificial Intelligence Approach*, R.S. Michalski, J.G. Carbonell, and T.M. Mitchell (eds.), Morgan Kaufmann, 463-482.

Schohn, G. and Cohn, D. (2000). Less is more: Active learning with support vector machines. In *Proceedings of the 17th International Conference on Machine-learning*, Stanford, CA, U.S.A., June 29-July 2, 2000, 839-846.

Sebastiani, F. (2002). Machine-learning in automated text categorization. *ACM Computing Surveys*, *34*(1), 1-47.

Semetko, H. A. and Valkenburg, P. M. (2000). Framing European politics: A content analysis of press and television news. *Journal of Communication*, *50*(2), 93-109.

Spertus, E. (1997). Somkey: Automatic recognition of hostile messages. In *Proceedings of the 5th International Conference on Intelligent User Interfaces*, Providence, RI, U.S.A., July 27-31, 1997, 1058-1065.

Sui, H., Khoo, C., and Chan, S. (2003). Sentiment classification of product reviews using SVM and Decision Tree induction. In *Proceedings of 13th ASIST SIG CR Workshop '2003*, Long Beach, California, October 18, 2003.

Tapanainen, P. and Järvinen, T. (1997). A non-projective dependency parser. In *Proceedings of the 5th Conference on Applied Natural Language Processing,* Washington, WA, U.S.A., March 31-April 3, 1997, 64-71.

Turney, P. D. (2002). 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, PA, U.S.A., July 7-12, 2002, 417-424.

Yang, Y. (1999). An evaluation of statistical approaches to text categorization. *Information Retrieval, 1*(1-2), 69-90.

Yang, Y. and Liu, X. (1999). A re-examination of text categorization methods. In *Proceedings of 22nd ACM International Conference on Research and Development in Information Retrieval*, Berkeley, CA, U.S.A., August 16-19, 1999, 42-49.