NORTHERN ARIZONA UNIVERSITY
The W. A. Franke College of Business

Sentiment Analysis on Chinese Web Forums using Elastic Nets: Features, Classification and Interpretation

Working Paper Series—11-11 | October 2011

Li Fan[1*], Yulei Zhang[2], and Hsinchun Chen[1]

[1]Department of Management Information Systems,

Eller College of Management

University of Arizona,

Tucson, Arizona, 85721

[2]The W. A. Franke College of Business,

Northern Arizona University,

Flagstaff, Arizona, 86011

E-mails: fanli@email.arizona.edu, yulei.zhang@nau.edu, hchen@eller.arizona.edu

* Corresponding Author

Sentiment Analysis on Chinese Web Forums using Elastic Nets: Features, Classification and Interpretation

1. Introduction

With the interactivity made possible by Web 2.0, a great deal of user-generated content about consumers' experiences and opinions, whether positive or negative, has been generated through online communities (Subrahmanian 2009). Content from these social media (e.g., blogs, web forums, online review sites, microblogging sites, etc.) can be readily extracted for further analysis (Cheong and Lee 2011). From the consumer's point of view, it is often helpful to refer to others' comments when making a purchase decision. On the other hand, for manufacturers, the sentiment expressed by people about their products is also an important basis for adjustments to product design, pricing, and marketing strategies: the growing availability and popularity of such opinion-rich resources create more opportunities to know what people are thinking about their products and services.

Sentiment analysis aims at helping people identify the opinions and emotions expressed by the authors in the online communities. Discovering sentiment is different from detecting factual statements which relies primarily on keyword-based scanning (Tweedie and Baayen 1998). Sentiment analysis requires complicated algorithms and more linguistic knowledge for understanding human language. Its essence is categorizing each text into two different polarities (positive and negative) based on features extracted from text. For both consumers and manufacturers, however, knowing only the sentiment polarity is not usually sufficient for decision-making, and additional information about why a certain product received a particular rating is far more helpful than just the rating itself. For example, both consumers and manufacturers would want to know why people like or dislike a particular product feature (e.g., people like the appearance of a car but dislike its safety). Many techniques have been applied to sentiment analysis. The semantic approaches are usually intuitive and easy to understand. However, they cannot always guarantee good accuracy. On the contrary, although the machine-learning based methods tend to have better accuracy, the complexity of these methods has led to difficulties in explaining how the polarity of a given text is determined. Therefore, a method which is able to classify each text into the right category and also provides meaningful interpretation for classification results would be preferable.

Over the past several decades, the Chinese speaking countries/regions (including mainland China, Taiwan, Hong Kong, Singapore and numerous other Chinese communities) have become more influential over the world's economy and politics (Ye et al. 2006). China not only has the world's largest population, but also a fast-growing market with huge numbers of consumers and manufacturers. According to the latest statistics of World Internet User, Chinese is the second most used language (followed by English) and the number of Chinese speaking people is now more than 20% of the world's population. Accordingly, Chinese sentiment analysis would enable organizations and individuals to investigate the Chinese market by analyzing consumer opinions. However, most previous sentiment analysis research has been conducted on English (Abbasi et al. 2008). There is much less research dealing with documents written in Chinese. Given the growing importance of the Chinese speaking market and the need to succeed in the global economy, it would become advantageous to understand the sentiments expressed by Chinese consumers.

Due to the rapidly increasing quantity of web sources and electronic texts in Chinese, Chinese language processing has received much attention in recent years. However, Chinese language processing is still quite challenging due to its unique linguistic characteristics (Wei et al. 2008). For example, unlike most Western languages, Chinese words do not have a boundary. In addition, Chinese has thousands of characters while English has only 26. Consequently, the features and feature-extraction techniques for Chinese text processing are different from those for English (Zheng et al. 2006).

Web forums, as a part of Web 2.0, are very influential online communication platforms for Chinese consumers wishing to exchange opinions (Das and Chen 2007). Web forums are a significant source of user-generated content which includes a broad range of emotional topics, valuable insights, and market sentiments. Unlike other forms of Web 2.0, Web forums have some unique characteristics such as a high volume of participants and large numbers of discussions. Compared with online review sites, forums are more discussion-driven and interactive (Stets 2006). There are also additional cues about participants' sentiments expressed in forums since sentiments are feelings which emerge from interactions. For the existing Chinese sentiment research, however, little effort has been put into the forum domain (Glance et al. 2008), while most papers focus on investigating opinions from online review sites (Ye et al. 2006; Tan et al. 2008). In this research we propose an interpretable and general framework for conducting sentiment analysis on Chinese Web forums.

The rest of the paper is organized as follows. In Section 2, we review the features used for classification and key classification techniques. Section 3 presents the proposed system framework design in detail. Section 4 describes the hypotheses and experiments. Section 5 discusses the selected features generated by Elastic Net. We conclude our study with discussions and suggestions for future research in Section 6.

2. Features and Classification Techniques

Feature representation is a critical element in sentiment classification, especially for online media. It is because online social media is very rich in interactive and stylistic information as compared to the traditional formal text (e.g., news articles, financial reports, etc.). Four major types of features have been adopted in previous literature on English sentiment analysis tasks in online social media: semantic features, stylistic features, syntactic features, and content-specific features. These features are described and explained below. Little previous research has adopted all of them into Chinese sentiment analysis research. Once a rich feature set has been extracted, another critical step in sentiment analysis is to build a classifier based on it.

2.1 Features

Semantic Features, also called word polarity features, are usually obtained based on pre-defined lexicons. A word has a positive score when it has good associations with positive items in the lexicon. Otherwise, it will be assigned a negative score. Although this type of features has been used a great deal in previous work (Ye et al. 2006), its classification accuracy largely depends on the quality of the sentiment lexicon. However, universally-accepted standard Chinese sentiment lexicon (such like SentiWordNet (http://sentiwordnet.isti.cnr.it/) in English) is still absent. Negation is another part of semantic features. In English, unlike traditional information retrieval, negation features play a more critical role in sentiment analysis. Similarly, negation is also an important part of features dealing with Chinese. Zagibalov incorporated a negation check by using two most commonly used Chinese negation words: 不(not) and 没(not) to improve the accuracy of sentiment classification(Zagibalov 2007).

Stylistic features, including lexical features and structural features (Zheng et al. 2006), provide evidence of people's writing styles, and have been applied to forum sentiment analysis in English and Arabic (Abbasi et al. 2008). Lexical features can be further divided into two types: character-level (e.g., number of digital character) and word-level (e.g., average word length). Structural features represent the layout of an article. For example, the number of sentences, number of paragraphs and whether one article contains a greeting or signature are all structural features.

Syntactic features include function words and punctuation marks. The discriminating capability of them has been shown in English text classification research (sentiment classification and authorship identification). It is because the usage of function words and punctuations determines syntactically how the author constructs a sentence (Pang and Lee 2008). Another type of syntactic feature is POS (part-of-speech) tags which are commonly leveraged in English sentiment analysis. POS can be considered as a form of word sense disambiguation (Pang and Lee 2008). For English sentiment classification, POS features include adjectives, nouns, adverbs and verbs.

Both stylistic and syntactic features can be referred to as ‘content-free’ features. Previous studies demonstrated that they are useful in English sentiment classification. Content-free features deserve special attention in our research because of the richness of such features in Web forums. They are able to reflect forum members’ writing styles and habits, which can be indicators of individual sentiment. For example, Abbasi et al. shows that positive movie reviews tend to be longer than negative ones (Abbasi et al. 2008). However, in a Chinese context, these two types of features have only been used in authorship identification (Zheng et al. 2008). Few studies have adopted them in Chinese sentiment analysis.

Content-Specific features, including character-level and word-level n-gram features, have been used much more often than content-free features. A character is the smallest processing unit in Chinese text. A character n-gram is a sequence of *n* consecutive Chinese characters. Although word boundaries do not exist in Chinese, a word which consists of multiple characters is the basic semantic unit in Chinese. Word n-grams consist of multiple Chinese words. Both of them have been applied in selected previous Chinese text classification and sentiment analysis work (Glance et al. 2008). Chinese text processing needs an additional segmentation step to break up the text into words (Zeng et al. 2011). In contrast, character-level features can be extracted in a simple way without any segmentation. In English, only word-level n-grams will be adopted because one English character is not necessarily meaningful. But the unique characteristic of Chinese language is that both word and character have semantic meaning. This makes it necessary to process features at both the character level and word level.

2.2 Classification Techniques

Semantic-Oriented approach is an unsupervised method that does not require prior training data. Specifically, it uses a sentiment lexicon to determine a degree of subjectivity via functions which count the number of positive and negative indicators (Pang and Lee 2008).This approach has been adopted by previous research (Tan et al. 2008; Tan and Zhang 2008; Zhang et al. 2009). The performance of the semantic-oriented approach is highly related to the quality of lexicon. Although training data with labeled information is not needed in the semantic-oriented approach and the result is easier to interpret, the performance may not be as good as Machine Learning approach.

Machine Learning approach (e.g., Naive Bayes and Support Vector Machine) has been shown to be powerful for sentiment analysis (Abbasi et al. 2008). Previous studies showed that the machine learning approach generally obtained better performance than the semantic-oriented approach in both English and Chinese sentiment analysis tasks. Support Vector Machine (SVM), as a statistical machine learning algorithm, has always been reported with the best performance for sentiment classification in both languages (Glance et al. 2008). But its mechanisms (e.g., kernel function and hyperplane) make it very difficult for people to interpret the results (Haasdonk 2005).

In general, the semantic-oriented approach can generate interpretable results, but its performance is not as good as a machine learning method (Stets 2006; Tan and Zhang 2008). Although machine learning methods have been shown to have excellent performance, it is very difficult to interpret the results and identify meaningful features.

2.3 Regularized Logistic Regression

Regularized Logistic Regression is a powerful and interpretable statistical model with good classification performance. It overcomes the deficiencies of machine learning approaches by improving interpretability and outperforms semantic-oriented approaches in terms of classification accuracy**.** In statistics, Logistic Regression is used to predict the probability of the occurrence of an event by fitting data to a logistic curve. Regularized Logistic Regression uses regularization which can avoid over-fitting. It has been shown recently to perform as well as SVM in terms of classification accuracy (Genkin et al. 2007). Ifrim et al. reported Regularized Logistic Regression achieved similar performance compared with SVM in both English and Chinese text classification tasks (Ifrim et al. 2008).

Regularized Logistic Regression has several advantages over SVM. In prior sentiment research using SVM, feature selection that can select a subset of 'good' features was always essential to improve the overall accuracy (Zou and Hastie 2005). But for Regularized Logistic Regression, its built-in automatic feature selection process is performed with classification simultaneously. Another advantage lies in its interpretability and simplicity. Regularized Logistic Regression is an interpretable classification model because it is able to assign a coefficient for each specific feature. Since different coefficients have different effects for the final classification result, we can distinguish the positive and negative features based on their coefficients.

There are three basic ways for regularization: Ridge, Lasso and Elastic Net (Zou and Hastie 2005). The difference among them is the form of regularization they adopt: Ridge takes L_2 penalty which minimizes the residual sum of squares of coefficients while Lasso imposes an L_1 penalty, Elastic Net adopts a linear combination of L_1 and L_2 penalty. Figure 1 shows the formula of Elastic Net with both L_1 and L_2 penalty (Zou and Hastie 2005). Compared to Ridge, Lasso is better because it is able to generate a sparse model with fewer features (Genkin et al. 2007). For many real-world classification tasks with a large number of features and few instances, however, Lasso is not an ideal method (Zou and Hastie 2005). Another type of Regularized Logistic Regression, Elastic Net, was proposed by Zou and Hastie in order to overcome the deficiency of Lasso (Zou and Hastie 2005). It is observed that Elastic Net outperforms Lasso when the number of features is much larger than the number of instances. In this work, we adopted Elastic Net since the number of n-gram features is usually very large, which caused the total number of features to greatly exceed the number of instances.

Figure 1: Elastic Net formula

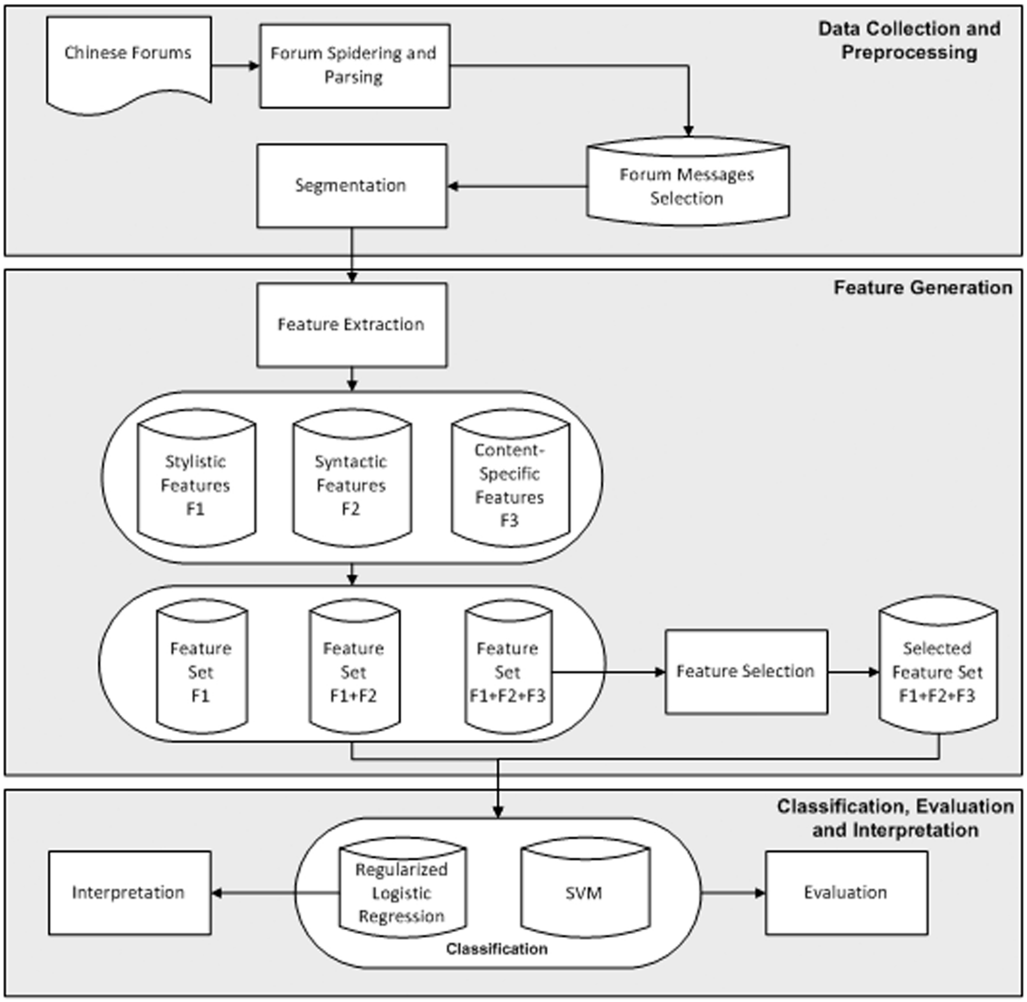
$$\hat{\beta}_{en} = (1+\lambda_2)\{\ \arg\min \| Y - \sum_{j=1}^{p} X_j\beta_j \|^2 + \lambda_2 \underbrace{\boxed{\sum_{j=1}^{p} |\beta_j|^2}}_{L_2\text{ penalty}} + \lambda_1 \underbrace{\boxed{\sum_{j=1}^{p} |\beta_j|}}_{L_1\text{ penalty}}\ \}$$

Classification performance and result interpretability are both very crucial for sentiment analysis. Most existing methods can not satisfy both requirements. To the best of our knowledge, none of the previous studies have adopted Elastic Net in Chinese sentiment analysis. Owing to the popularity of SVM in prior research, we conducted a performance comparison between SVM and Regularized Logistic Regression. Furthermore, we also demonstrated the interpretability of Regularized Logistic Regression in sentiment analysis for Chinese forums.

3. System Design

Figure 2 illustrates our proposed framework. It consists of three major phases: data collection and preprocessing, feature generation, and classification, evaluation and interpretation.

Figure 2: Chinese forum sentiment analysis framework



3.1 Data Collection

In the data collection phase, a spidering program has been developed to collect HTML pages from Chinese forums. Parsing programs are developed to parse out the messages from raw HTML pages and store them in the database. After the parsing step, a subset of parsed messages will then be selected into our testbed. We adopted *Stanford Chinese Word Segmenter* (http://nlp.stanford.edu/software/segmenter.shtml) for word segmentation in our framework. This tool is based on Conditional Random Field and can achieve excellent segmentation performance.

To evaluate our Chinese forum sentiment analysis framework, we chose two famous forums from different domains as our testbed. The first testbed was selected from the Chinese Yahoo Stock forum (http://mystock.cn.yahoo.com/guba/index.html) in which the topics relate to the Shanghai Stock Index. The second testbed was obtained from a forum belonging to one of the largest Chinese digital product websites (http://bbs.zol.com.cn). The discussion board we chose is about DELL laptops. Both forums are all very influential in their own domain with a large number of discussions every day.

The Yahoo Stock forum has prediction indicators ('up' or 'down') which are provided by posters in every single message. With regards to the stock exchange, bullish opinions (up) can be considered as positive sentiment while bearish opinions (down) can be treated as negative sentiment, allowing us to use them as our human-generated sentiment tags for each message. The DELL laptops forum does not have such explicit user-generated tags, so we asked a Chinese domain expert to manually tag each posting as positive or negative. In the Stock forum, we collected 580 messages consisting of 290 bullish ones and 290 bearish ones as our first testbed. We also randomly selected 500 messages from the Laptops forum with 250 positive and 250 negative messages as our second testbed.

3.2 Feature Generation

Feature generation process has been used in several previous research works in order to find out the features which can best represent candidate texts (Wei et al. 2008; Zheng et al. 2007; Zhai et al. 2011). In other words, in this research, it generates text features that result in the best classification performance for the postings from our forum collection. Because semantic features are usually adopted by a semantic-oriented approach, we did not include them in the feature set. Finally, three types of features, i.e., stylistic features (F1) (e.g. Number of Characters, Number of Sentence, Avg. Number of Word / Paragraph), syntactic features (F2) (e.g. Frequency of punctuation, Frequency of function words) and content-specific features (F3) are adopted. The content-free features (42 stylistic features and 87 syntactic features) we utilized are adopted from previous research for authorship analysis in Chinese forums (Zheng et al. 2007). Since the number of content-specific features is dependent on specific application domains, high-frequency n-grams were extracted from corpus as features (Zheng et al. 2007; Pang and Lee 2008). Finally, 19,134 features for the Stock dataset and 7,334 features for the Laptops dataset were used. Due to the unique characteristics of the Chinese language, we can extract features from both character-level and word-level N-grams (unigram, bigram and trigram). Three feature sets were generated incrementally based on different combinations of feature types:

- Stylistic Features (F1)
- Stylistic + Syntactic Features (F1+F2)
- Stylistic + Syntactic + Character N-gram + Word N-gram Features (F1+F2+F3)

Feature selection is an effective way to enhance the classification performance, and Information Gain (IG) has been reported as an effective feature selection method (Abbasi et al. 2008). Therefore we also constructed an additional feature set (selected F1+F2+F3) with features selected by IG. Elastic Net does not need to do feature selection because it can automatically use the top features with high discriminating power to build its classification model (Zou and Hastie 2005). So no feature selection is applied to Elastic Net.

3.3 Classification, Evaluation and Interpretation

SVM and Elastic Net are the two classification techniques adopted in our research. There are several other machine learning approaches available for sentiment analysis (e.g. Naive Bayes and Maximum-entropy-based method). In this study, we choose SVM because it has been consistently reported to have the best performance. In our experiment, we used SVM implemented in WEKA data mining package (http://www.cs.waikato.ac.nz/~ml/weka) which is a widely used data mining tool. Elastic Net is implemented by using the glmnet package (http://cran.r-project.org/web/packages/glmnet/index.html) in R statistics software (www.r-project.org). To compare their performance, we built two classifiers and adopted two criteria for evaluation: Accuracy and F-measure, both of which have been widely used in previous research (Wei et al 2008; Zheng et al. 2007).

Elastic Net enables the interpretation of polarity and the significance of different features based on coefficients. In Elastic Net, features with positive coefficients have a positive impact on the prediction value, while features with negative coefficients play a negative role. In the interpretation phase, we extracted top-ranked features ordered by their coefficients generated by the model. Furthermore, we also compared the selected feature sets generated by Elastic Net and Information Gain to figure out their differences.

4. Hypotheses and Experiments

We developed two groups of hypotheses to examine the performance of different feature sets and techniques in Chinese forum sentiment analysis.

Hypothesis 1: By adding more feature types, Chinese sentiment classification performance will improve.

> *Hypothesis 1.1*: Combining all content-free features can outperform the use of stylistic features alone in Chinese sentiment classification, i.e., F1+F2 > F1
>
> *Hypothesis 1.2*: Combining content-free features with content-specific features can achieve better results than using only content-free features, i.e., F1+F2+F3 > F1+F2

Hypothesis 2: Regularized Logistic Regression can achieve similar performance results as compared to SVM.

Table 1 shows the average Accuracy and F-measure of each dataset after classification with SVM and Elastic Net. 10-fold cross validation was used to evaluate both classifiers. As shown in Table 1, for the first three feature sets, classification accuracy and F-measure increase as additional feature types are added. The results also show that Elastic Net's performance is comparable to SVM using any feature set.

Table 1: Accuracy and F-measure of both datasets

Dataset	**Feature Sets**	**Accuracy**		**F-measure**	
		SVM	Elastic Net	SVM	Elastic Net
Stock	F1	70.4%	69.6%	71.2%	70.7%
	F1+F2	74.9%	73.8%	75.4%	73.4%
	F1+F2+F3	80.8%	84.3%	80.2%	84.4%
	Selected F1+F2+F3	82.3%	-	82.7%	-
Laptops	F1	61.1%	61.4%	61.0%	61.0%
	F1+F2	69.5%	68.5%	69.5%	68.2%
	F1+F2+F3	84.1%	86.6%	84.6%	86.7%
	Selected F1+F2+F3	86.0%	-	86.2%	-

In order to determine the statistical significance of Hypothesis 1, a pairwise t-test was conducted. The results shown in Table 2 suggest that both Hypothesis 1.1 and Hypothesis 1.2 are supported (all p values < 0.0001).

Table 2: T-test results of H1.1 and H1.2

		SVM		**Elastic Net**	
	Hypotheses	P value on accuracy	P value on F-measure	P value on accuracy	P value on F-measure
Stock	H1.1 F1+F2>F1	<0.0001**	<0.0001**	<0.0001**	<0.0001**
	H1.2 F1+F2+F3>F1+F2	<0.0001**	<0.0001**	<0.0001**	<0.0001**
Laptops	H1.1 F1+F2>F1	<0.0001**	<0.0001**	<0.0001**	<0.0001**
	H1.2 F1+F2+F3>F1+F2	<0.0001**	<0.0001**	<0.0001**	<0.0001**

We also conducted another pair-wise t-test to compare the discriminating power of Elastic Net and SVM. In this test, we adopted the best results generated by both. Specifically, in SVM we chose selected F1+F2+F3 and in Elastic Net we chose F1+F2+F3. From the testing results (shown in Table 3) we find that the classification accuracy and F-measure are similar between using Elastic Net without feature selection and using SVM with the selected feature set. There is no significant difference between the performances of the two techniques (all p values greater than 0.05). Thus, Hypothesis 2 is supported.

Table 3: T-test results of H2

	Stock		**Laptops**	
Comparison	P value on accuracy	P value on F-measure	P value on accuracy	P value on F-measure
SVM > Logistic Regression OR SVM < Logistic Regression	0.2056	0.2546	0.3658	0.3045

Therefore, the experiment results show that the performance of SVM and Regularized Logistic Regression is comparable. This result is consistent with previous text categorization research (Ifrim et al. 2008; Zhang and Oles 2001). Our study indicates the similar finding in Chinese sentiment analysis. With a performance similar to that of SVM, the advantage of Elastic Net lies in its ease of interpretability, which will be discussed in the following section.

5. Feature Interpretation

In the feature interpretation phase, we first built the Elastic Net model using the Stock and Laptops datasets. Then, we ranked all the features based on coefficients generated by Elastic Net. Finally, we selected some examples of top-ranked positive and negative features from each testbed. The selected ones with feature name, coefficient and feature type are shown in the Table 4.

Table 4: Top positive and negative features in Stock and Laptops dataset

Dataset	Positive Features			Negative Features		
Stock	Coef.	Feature	Feature Type	Coef.	Feature	Feature Type
	0.458	指数_看_涨 (index will go up)	F3 (W) (Trigram)	-1.877	Simpson’D	F1
	0.355	大盘_涨幅 (amount of increase)	F3 (W) (Bigram)	-0.925	清仓 (close out)	F3 (W) (Unigram)
	0.295	较_佳 (relatively good)	F3 (C) (Bigram)	-0.506	很_弱 (very weak)	F3 (C) (Bigram)
	0.198	良好 (excellent)	F3 (W) (Unigram)	-0.490	业绩_下滑 (performance decreases)	F3 (W) (Bigram)
	0.108	四万亿 (4000 billion)	F3 (W) (Unigram)	-0.250	没有_利_好 (no bullish news)	F3 (W) (Trigram)
Laptops	Coef.	Feature	Feature Type	Coef.	Feature	Feature Type
	0.798	重要_的 (important)	F3 (W) (Bigram)	-0.897	缺陷 (disadvantage)	F3 (W) (Unigram)
	0.704	很_满意 (very satisfied)	F3 (W) (Bigram)	-0.501	不_怎么样 (not very good)	F3 (W) (Bigram)
	0.530	合理 (reasonable)	F3 (W) (Unigram)	-0.342	有点_贵 (a little expensive)	F3 (W) (Bigram)
	0.462	镁_合金 (magnesium alloy)	F3 (W) (Bigram)	-0.290	响 (noisy)	F3 (C) (Unigram)
	0.389	不错 (nice)	F3 (W) (Unigram)	-0.137	缝隙 (seam)	F3 (W) (Unigram)

Note: W = word-level features, C = Character-level features

Table 4 shows that most top-ranked features have strong sentiment polarity. Features that have a positive coefficient value are the ones with positive sentiment and features with negative coefficient tend to contain negative sentiment.

In the ‘Feature Type’ column we find most top-ranked features are content-specific features (F3). The diversity of F3 implies the necessity of selecting N-gram features from both word-level and character-level. Content-specific features that contain domain knowledge provide good discriminating power. Some features, as the characteristics of certain products, can be used to identify consumers’ opinion in a more detailed way. For instance, the feature *镁_合金 (magnesium alloy)*, from the Chinese laptops dataset, is a positive feature. This material has been widely used in Dell laptops to provide more impact strength on the laptop body. It suggests that Chinese consumers are more likely to give positive feedback to the usage of this material. Some other features, although not part of a product, can also influence the product’s performance in an indirect way. For example, the feature *四万亿 (4000 billion)* from the Chinese stock dataset, is also a positive feature. It is the amount of the rescue package proposed by the Chinese government in 2008 in response to the economic recession. It indicates that Chinese investors are confident about the stock market when they talk about this rescue package.

Additionally, our results show that some content-free features are also valuable. One example here is the Simpson’s D stylish feature -- a vocabulary richness measure defined by Simpson (Tweedie and Baayen 1998). This measure has a very high negative coefficient (-1.877) from the Elastic Net in stock dataset. It implies that, for a posting with a fixed length, people who use a small set of certain frequently-repeated words are more likely to convey negative sentiment than those who use a large set of less frequently-used words. For example, if one investor holds a negative attitude towards the stock market, he/she may complain about the market and will repeat a few words many times, words such as ‘decrease’ or ‘bearish’. On the other hand, investors with positive sentiments tend to express their predictions in many different ways, which means a greater number of less frequently-used words will be used.

To compare the features selected by Elastic Net and IG, we examined features (before and after feature selection) from all feature types in both datasets. Selection Ratio is defined as the percentage of selected features to the features in the initial feature set. The results indicate that the selection ratio of Elastic Net is no larger than that of IG for all features. This implies that Elastic Net is able to achieve identical classification performance using often fewer features compared with IG. Moreover, although the number of content-specific features is much larger than the number of content-free features, the selection ratio of content-specific features is not necessarily larger than the content-free features. This may suggest that the role of content-free features in sentiment analysis is also significant.

6. Conclusion

The continuing growth in global products and services underscores the significance of understanding the opinions and sentiments of international consumers using languages other than English. China, as a nation with the largest population in the world, has been chosen by many enterprises as their key emerging market. To address the issue of accurately detecting Chinese consumers’ opinions, we proposed a framework to analyze sentiment from Chinese Web forums. We adopted a Regularized Logistic Regression-based classification method to categorize online sentiment based on comprehensive language features. We were able to obtain meaningful classification results that may have significant market implications. Experimental studies conducted on Chinese Stock and Laptops forums demonstrated good performance of our proposed framework in terms of accuracy and interpretability.

Currently, a couple of interesting issues in this research area are still open. In addition to the textual features, other non-textual features such as link features (e.g., reply-to relationship or agree-or-disagree relationship between forum postings) could be leveraged in our framework. Because this sentiment research is performed at the document level, it would be a good extension to study Chinese sentiment at the paragraph or phrase level. Moreover, we would be interested in applying this framework for other languages by considering their language characteristics.

Acknowledgement

This work is supported by the NSF Computer and Network Systems (CNS) Program, “(CRI: CRD) Developing a Dark Web Collection and Infrastructure for Computational and Social Sciences,” (CNS-0709338), September 2007 – August 2010.

References

Abbasi, A., Chen, H. & Salem, A. (2008). Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums. *ACM Transactions on Information Systems (TOIS),* 26(3), pp. 1-34.

Cheong, C. & Lee, V. (2011). A microblogging-based approach to terrorism informatics: Exploration and chronicling civilian sentiment and response to terrorism events via Twitter. *Information System Frontier*, 13, 45–49

Das, S. R. & Chen, M. Y. (2007). Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web. *Management Science*, 53(9), pp. 1357-1388.

Genkin, A., Lewis, D.D. & Madigan, D. (2007). Large-Scale Bayesian Logistic Regression for Text Categorization. *Technometrics*, 49(3), pp.291-305.

Glance, N., Hurst, M. & Nigam, K. (2008). Deriving Marketing Intelligence from Online Discussion. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'08),* pp.419-428

Haasdonk, B. (2005). Feature Space Interpretation of SVMs with Indefinite Kernels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(4), pp. 482-492.

Ifrim, G., Bakir, G. & Weikum, G. (2008). Fast Logistic Regression for Text Categorization with Variable-Length N-grams. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'08),* pp.354-362.

Li, J., Zheng, R. & Chen, H. (2006). From Fingerprint to Writeprint. *Communications of ACM*, 49(4), pp.76 - 82.

Li, L. & Sun, M. (2007). Experimental Study on Sentiment Classification of Chinese Review using Machine Learning Techniques". *International Conference on Natural Language Processing and Knowledge Engineering,* pp.393-400.

Pang, B. & Lee, L. (2008). *Opinion Mining and Sentiment Analysis*. Now Publishers Inc.

Stets, J. E. (2006). Emotions and Sentiments. *Handbook of Social Psychology*, J. Delamater, ed., pp. 309-355: Springer US

Subrahmanian, V.S. (2009). Mining Online Opinions. *Computer*, pp. 88 – 90.

Tan, S., Wang, Y. & Cheng, X. (2008). Combining Learn-based and Lexicon-based Techniques for Sentiment Detection without Using Labeled Examples. *ACM Conference on Research and Development in Information Retrieval (SIGIR'08),* pp. 743-744.

Tan, S. & Zhang, J. (2008). An empirical study of sentiment analysis for Chinese documents. *Expert Systems with Applications*, 34(4), pp. 2622-2629.

Tweedie, F.J. & Baayen, R.H. (1998). How Variable May a Constant Be? Measures of Lexical Richness in Perspective. *Computers and the Humanities*, 32(5), pp.323–352.

Wei, Z., Miao, D. & Chauchat, J.H. (2008). Feature Selection on Chinese Text Classification Using Character N-Grams. *The 3rd International Conference on Rough Sets and Knowledge Technology,* pp.500-507.

Ye, Q., Shi, W. & Li, Y. (2006). Sentiment Classification for Movie Reviews in Chinese by Improved Semantic Oriented Approach. *Proceedings of the 39th Hawaii International Conference on System Sciences (HICSS'06)*

Zagibalov, T. (2007). Kinds of Features for Chinese Opinionated Information Retrieval. *45th Annual Meeting of the Association for Computational Linguistics(ACL'07)*, pp. 37-42

Zeng, D., Wei, D., Chau, M. & Wang, F. (2011). Domain-specific Chinese word segmentation using suffix tree and mutual information. *Information System Frontier*, 13, 115–125

Zhai, Z., Xu, H. & Kang, B. (2011). Exploiting Effective Features for Chinese Sentiment Classification. *Expert Systems with Applications,* 38(8), pp. 9139-9146

Zhang, C., Zeng, D. & Li, J. (2009). Sentiment analysis of Chinese documents: From sentence to document level. *Journal of the American Society for Information Science and Technology (JASIST)*, 60(12). pp.2474-2487.

Zhang, T. & Oles, F.J. (2001). Text Categorization Based on Regularized Linear Classification Methods. *Information Retrieval*, 4(1), pp.5-31.

Zheng, R., Li, J., Chen, H. & Huang, Z. (2006). A Framework for Authorship Identification of Online Messages: Writing-Style Features and Classification Techniques. *Journal of the American Society for Information Science and Technology (JASIST)*, 57(3), pp.378-393.

Zou, H. & Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society*, 67(2), pp.301-320.