2012 IEEE International Conference on Systems, Man, and Cybernetics
October 14-17, 2012, COEX, Seoul, Korea

Statistical Text Analysis and Sentiment Classification in Social Media

Sang-Hyun Cho
Dept. of Computer Engineering
The Catholic university of Korea
Bucheon, Republic of Korea
cshgreat@catholic.ac.kr

Hang-Bong Kang
Dept. of Digital Media
The Catholic University of Korea
Bucheon, Republic of Korea
hbkang@catholic.ac.kr

Abstract— **In this paper, we propose a new method of classifying tendencies and opinions in texts of multiple sentence length extracted from social media and covering both formal and informal vocabularies. To extract contextual information from the texts, we carry out computations based on keywords, the position of the sentence and the flow of sentiments in the multiple texts. A feature vector for the given text is constructed from the contextual information, and is then classified with a Support Vector Machine (SVM) classifier as positive, negative or neutral. Our method performs well in classifying the gradient of sentiments expressed in social media.**

Keywords- social media; text sentiment; SVM

I. Introduction

Recently, social network services (SNS) such as Twitter and Facebook have become popular. Finding sentiments and tendencies among the opinions stated in tweets or texts within SNS connected to consumer products can be very important in planning marketing strategies. Essentially, the overall sentiment in a text is determined by the semantics of each sentence and the contextual information. Therefore, it is essential to extract semantic and contextual information in order to understand and classify the overall sentiment of the writer correctly.

To perform opinion classification, various methods have been used. There have been studies that simultaneously use rule-based methods and statistics-based methods [1-3]. Various rules and statistics are utilized in order to consider adjectives as opinion words and to determine the semantic orientation of the adjectives. In addition, studies of document-level analysis use learning algorithms such as Naïve Bayes, Maximum Entropy (ME) classification and SVM [4-5]. For machine learning, handwork was performed in order to generate the bulk of the tagged corpus.

The earliest work of automatic sentiment classification at document-level is in [6]. Here, several machine learning approaches are combined with common text features to classify movie reviews from the Internet Movie DataBase(IMDB). Dave et al. [7] designed a classifier based on information retrieval techniques for feature extraction and scoring. Mullen and Collier integrated PMI values [8], Osgood semantic factors and some syntactic relations into the features of SVM [9]. Yang et al. [10] used web-blog to construct corpora for sentiment analysis and emotion icons assigned to blog posts as indicators of users' moods. They used SVM and CRF learners to classify sentiments at the sentence level and then investigated several strategies to determine the overall sentiment of the document. Read et al. [11] used emoticons to form a training set for sentiment classification. They collected texts from Usenet newsgroup and divided them into "positive" and "negative" samples. An emoticon is classified by a trained SVM and a Naïve Bayes classifier. As a result, they obtain up to 70% of accuracy on the test set. Go et al. [12] used Twitter to collect training data and then perform a sentiment classification. They also construct corpora by using emoticons to obtain "positive" and "negative" samples, and then use various classifiers.

To deal with semantics from texts, it is necessary to build appropriate dictionaries. However, most of traditional sentiment analysis methods are performed by only using the formal dictionary [6]. This strategy is not effective for this type of classification because internet sentence such as tweets in Twitter usually contain informal vocabulary such as emoticons and newly coined words. Emoticons are widely used to express the emotion of the internet users. It is difficult to classify the emotions behind the emoticons because the structure in emoticons is irregular. Newly coined words are also popular in the SNS. In addition, the same word may have different meanings depending on the context. Therefore, it is necessary to build sentiment-based domain dictionaries.

For the contextual information analysis from multiple texts, it is necessary to extract key words for evaluating the importance of the sentence. The position of the sentence in multiple texts is also to be considered because the first or last sentences usually play an important role in determining the sentiment behind the entire text. In addition, the flow of sentiments generated from each sentence gives a clue in determining the sentiment among multiple texts.

In this paper, we propose a new approach to text sentiment classification using contextual information specifically obtained from paragraphs. To classify the sentiments of paragraph-length texts in SNS, we compute contextual information based on domain-based keywords, the position of the sentence, and the flow of sentiments. The rest of this paper

978-1-4673-1714-6/12/$31.00 ©2012 IEEE

is organized as follows. In Section II, we discuss text sentiment classification method using sentiment-based domain dictionary. Section III describes how to extract contextual information in the paragraph. Section IV shows our experimental results.

II. Text Sentiment Classification Using Sentiment-Based Domain Dictionary

Fig. 1 shows an overview of our approach. First, we perform a sentence sentiment classification using sentiment-based domain dictionaries covering formal and informal vocabularies. Then, we extract contextual information from keywords, the position of the sentence and the flow of sentiments in a paragraph. Finally, we classify the overall sentiment in the paragraph by computing a contextual information degree, which is the linearly combined weighted sum of contextual information.

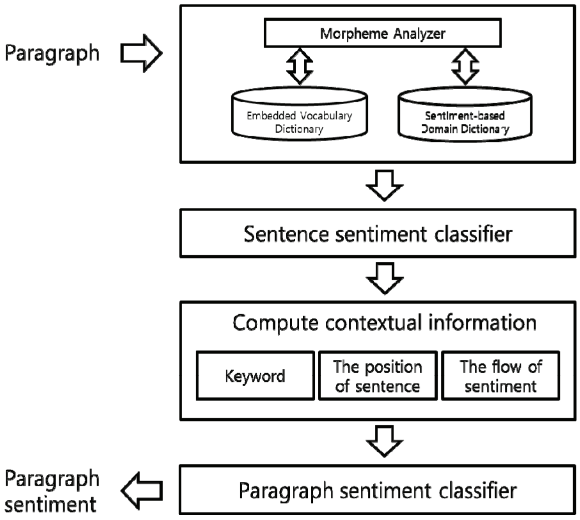

Paragraph
Morpheme Analyzer
Embedded Vocabulary Dictionary
Sentiment-based Domain Dictionary
Sentence sentiment classifier
Compute contextual information
Keyword
The position of sentence
The flow of sentiment
Paragraph sentiment
Paragraph sentiment classifier

Figure 1. Overview of our method

To classify the sentiment in each sentence, we use a sentiment classification method, which is presented in [13]. The sentiment classification method uses sentiment-based domain dictionaries covering both formal and informal vocabularies, since the same word may have different meanings depending on the context, and newly coined words and Emoticons are widely used in SNS.

We constructed five sentiment-based domain dictionaries as shown Table I. Various internet text resources such as relevant review webpages, personal blogs and tweets that were searched by relevant keywords such as 'iPhone' in the consumer product domain are explored. However, the number of obtained sentiment features using this method was too small to construct a sentiment dictionary. To expand our emotion-oriented vocabulary, we use an internet thesaurus which is provided by NAVER [5] in Korea. The sentiment weight of each word in our dictionary is calculated by the TF-IDF (term frequency - inverse document frequency) method since some words that frequently appear in the specific sentences provide useful information how to classify the sentiment of sentence as similar previous researches[1,6,13] suggested.

The sentiment of sentence is determined as follows: each sentence of input paragraph-length text is decomposed into five morpheme classes (noun, verb, adjective, adverb and emoticon) by the morpheme analyzer. An input sentence is represented as a feature vector from decomposed morphemes classes. Each component of the feature vector has a sentiment weight determined from the sentiment based domain dictionaries. The feature vector is classified into one of three sentiments (positive, negative, neutral) by sentence sentiment classifier. More details are in [13].

TABLE I. Our Sentiment-based Domain Dictionary

domain	Sentiment	noun	verb	adjec-tive	ad-verb	emo-ticon
consumer product	positive	261	100	79	42	19
	negative	359	230	84		18
Person	positive	237	92	75		19
	negative	321	214	81		18
Travel	positive	258	104	85		19
	negative	342	220	83		18
Food	positive	268	109	82		19
	negative	344	226	80		18
Movie	positive	257	96	77		19
	negative	338	224	78		18

Untypical emoticons are frequently shown in tweets. To deal with untypical emoticons, we convert untypical emoticons into reference emoticons using Bayesian framework. To compute probability distribution of a given emoticon, we use emoticon components like Table II from our sentiment-based domain dictionaries.

TABLE II. Our Emoticon Component

^	=	_	-	.
;	+	o	@	~
3	O	▽	♡	♥
★	,	$	(	)
/	>	<	。	‾
ㄷ	#	?	:	$
%	&	*	ㅁ	ㅇ
ㅅ	ㅋ	ㅎ	ㅜ	ㅠ

To convert an untypical emoticon to a reference emoticon, we use a maximum likelihood method. Let q be an untypical emoticon and q_{ref} a reference emoticon, then the maximum likelihood is

$$p(q_{ref} \mid q) \propto p(q \mid q_{ref})(q_{ref}) \qquad (1)$$

Using maximum likelihood method, we choose the optimal reference emoticon [13]. For example, untypical emoticon '^_____^' is converted to the reference emoticon '^_^'.

III. Contextual Information in a Paragraph

After classifying the sentiment of each sentence, we compute the contextual information of the entire paragraph. To do this, we extract reliable keywords from the domain training data. The weight of importance for each keyword is based on the frequency rate in our training data set. We normalize the frequency value of each word with a weighting factor between zero and one. Then sentence weight is calculated as follows.

For each sentence in the text,

$$\alpha(S) = \frac{1}{K}\sum_{i} \omega_i k_i(S) \qquad (2)$$

, where S is input sentence, $k_i(S)$ is containing i th keyword in sentence S , $\omega_i \in$ R is its weight and K is the number of keywords in sentence S .

The position of the sentence is also important, since users express their main opinions or sentiments with the opening and / or closing sentence of their texts. In particular, in many cases, SNS users express main opinion at the opening sentence due to limited space in SNS. To reflect this fact, we compute the position of sentence in a paragraph as follows.

$$\beta(S) = \begin{cases} \frac{1}{e} e^{-\mathrm{idx}(S)} \\ \frac{1}{e} e^{\mathrm{idx}(S)} \end{cases} \quad \text{where } -1 \le \mathrm{idx}(S) \le 1 \qquad (3)$$

, where $\mathrm{idx}(S)$ is a normalized index of sentence S when the middle index of sentence is zero.

In addition, the flow of sentiment among texts is useful in computing contextual information. For example, consider the following four sentences.

- This smart phone has more functions than others.
- The design looks good.
- The battery lasts longer than other ones.
- But this smart phone is too expensive to buy.

The first three sentences have positive sentiments and the last one has negative, thereby an overall sentiment of above texts is positive if we use a simple voting classification method. However, a sudden change of sentiment as in the above example can determine the overall sentiment of the entire paragraph.

The flow of sentiment of sentences is represented as follow.

$$\gamma(S) = \begin{cases} 1 \quad \text{if sentiment is not changed} \\ e^{nps(S)} \ \text{if sentiment is changed} \end{cases} \qquad (4)$$

, where $nps(S)$ is a number of previous consecutive sentences which have the same sentiments.

From (1) - (3), the degree of the contextual information for each sentence S is computed as follows:

$$C(S) = \omega_1\alpha(S) + \omega_2\beta(S) + \omega_3\gamma(S) \qquad (5)$$

where ω_1, ω_2 and ω_3 are weight factors. We set ω_1 = 0.4, ω_2 = 0.3 and ω_3 = 0.3. These factors are empirically determined.

For contextual information of each sentence in given paragraph, we construct a feature vector to determine the sentiment of the paragraph. The feature vector is constructed from *m* sentences with high contextual information degree. In our approach, we set *m* to 4 sentences to construct a feature vector. We classify the feature vector of paragraph into 3 sentiments (positive, negative and neutral) using SVM-based paragraph sentiment classifier.

IV. Experimental Results

We use KLT (Korean Language Technology) to analyze the morphemes of a sentence. SVM is trained with a linear kernel function. To construct a test data set, we collect texts corresponding to 'food', 'travel', 'movies' and 'consumer products' contexts from SNS such as Twitter, Facebook and Me2Day. We used two user accounts for each SNS and collected various sentences with respect to relevant keywords. Table III shows our test data set. The sentiment based domain dictionary is manually selected to test each domain data.

Our test data set is divided into ten equal parts; we execute 10-fold cross validation. We construct four domain-based sentiment dictionaries where each vocabulary is associated with a sentiment and its weight. A word is associated with a sentiment if at least 7 of 10 students agree to it.

To evaluate our method, we use an F1-measure with precision and recall, which is widely used in the retrieval and classification of documents.

$$\text{precision}(\rho) = \frac{\text{Number of correctly classified sentence correspond to emotion}}{\text{Number of all classified sentence correspond to emotion}} \qquad (6)$$

$$\text{recall}(\gamma) = \frac{\text{Number of correctly classified sentence correspond to emotion}}{\text{Number of all sentence correspond to emotion}} \qquad (7)$$

$$F_1 - \text{measure} = \frac{2\gamma\rho}{\gamma + \rho} \qquad (8)$$

TABLE III. ORGANIZATION OF OUR TEST DATA SET

domain	Emoticon exist	no	yes	no	yes	Total
	Newly coined words exist	no	no	yes	yes	
consumer product	Positive	180	31	6	46	263
	Negative	110	9	1	2	122
	Neutral	517	37	17	45	616
Travel	Positive	145	76	45	21	287
	Negative	33	8	32	9	82
	Neutral	71	18	6	2	97
Food	Positive	86	56	36	4	182
	Negative	13	6	3	1	23
	Neutral	268	15	16	3	302
Movie	Positive	316	127	100	40	583
	Negative	68	13	8	3	92
	Neutral	87	2	9	1	99
Total		1,894	398	279	177	2,748

Table IV shows the result of computing contextual information from our test data. Our method has better performance than other methods in most domains.

Case 1 : Only keywords are used.

Case 2 : Only the position of the sentence is used.

Case 3 : Only the flow of sentiment is used.

Case 4 : Our proposed method.

We compare our sentiment classification result with Zhuang et al.'s approach [14] in the 4-domain (consumer product, travel, food and movie). Zhuang's approach provides multi-knowledge based approach integrating keyword, statistical analysis and specific domain knowledge. Note that we modified Zhuang's approach to apply it to the Korean paragraph. To compare two methods, we used feature keywords and opinion keywords as our keywords and sentiment dictionaries. Fig. 2 shows F1-measure comparison results. Our approach has a much higher F1-measure on test data than Zhuang's approach.

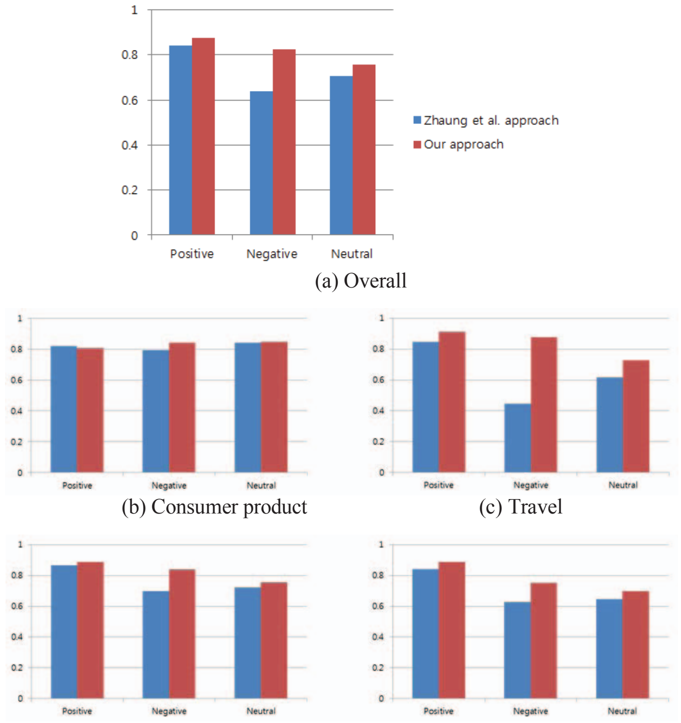
The first reason is that they omitted to consider the relationship between sentences. In many cases, the keyword-based, feature-opinion pair is invalid due to the complexity of the sentences. The second reason is that they omit to consider informal vocabulary such as newly coined words and emoticon. The informal vocabulary in social media is essential in determining the correct sentiment of paragraph.

However, in the consumer product domain, our approach has a lower F1-measure on test data than Zhuang's for positive sentiment. This is due to the absence of keywords in major opinion sentences.

V. CONCLUSION

In this paper, we propose a new approach to sentiment classification at paragraph length using contextual information and sentiment-based domain dictionaries covering formal and informal vocabularies. Contextual information such as key words, the position of the sentence, and the flow of sentiment are computed in texts of multiple sentence length. A feature vector for a given text is constructed from the contextual information and is then classified by the Support Vector Machine (SVM) classifier as positive, negative or neutral. Our method performs well in classifying the sentiments expressed in the multiple texts of social media.

In the future work, we will further improve our approach to extract users' gender, age and preferences from texts. In addition, we will research an on-line learning method to take account more quickly of newly created emoticons and coined words.



(a) Overall

(b) Consumer product

(c) Travel

(d) Food

(e) Movie

Figure 2. Comparison results.

Acknowledgements

This research is supported by Ministry of Culture, Sports and Tourism (MCST) and Korea Culture Content Agency (KOCCA) in the Culture Technology (CT) Research & Development Program 2011 and is also supported by the Basic Science Research Program through the National Research Foundation (NRF) funded by the Ministry of Education, Science and Technology (2010-0024641).

References

[1] Pimwadee Chaovalit and Lina Zhou. Movie review mining: A comparison between supervised and unsupervised classication approaches. In Proceedings of HICSS 2005, vol.4. 2005.

[2] A. Esuli and F. Sebastiani, “Determining term subjectivity and term orientation for opinion mining,” in Proceedings of the European Chapter of the Association for Computational Linguistics (EACL), 2006.

[3] K. Lerman, S. Blair-Goldensohn, and R. McDonald, “Sentiment summarization: Evaluating and learning user preferences,” In Proceedings of the European Chapter of the Association for Computational Linguistics (EACL), 2009.

TABLE IV. MULTIPLE TEXTS TEST RESULT

domain	Case	emotion	ρ	γ	F_1
Overall	Case 1	Positive	0.8256	0.8881	0.8519
		Negative	0.7196	0.6234	0.6660
		Neutral	0.6883	0.6822	0.6830
	Case 2	Positive	0.8366	0.8860	0.8602
		Negative	0.8302	0.7354	0.7755
		Neutral	0.7425	0.6966	0.7158
	Case 3	Positive	0.8046	0.8904	0.8437
		Negative	0.8382	0.7299	0.7755
		Neutral	0.7129	0.6549	0.6790
	Case 4	Positive	0.8330	0.9202	0.8734
		Negative	0.9159	0.7554	0.8262
		Neutral	0.7851	0.7345	0.7563
consumer product	Case 1	Positive	0.7105	0.9642	0.8181
		Negative	0.9545	0.7500	0.8399
		Neutral	0.8636	0.7307	0.7916
	Case 2	Positive	0.7500	0.8571	0.8000
		Negative	0.8148	0.8461	0.8301
		Neutral	0.9130	0.7500	0.8235
	Case 3	Positive	0.6944	0.8928	0.7812
		Negative	0.8750	0.7500	0.8076
		Neutral	0.8636	0.7307	0.7916
	Case 4	Positive	0.7352	0.8928	0.8064
		Negative	0.9545	0.7500	0.8399
		Neutral	0.8461	0.8461	0.8461
Travel	Case 1	Positive	0.8627	0.8301	0.8461
		Negative	0.4444	0.4444	0.4444
		Neutral	0.5925	0.6400	0.6153
	Case 2	Positive	0.8727	0.9056	0.8888
		Negative	0.8571	0.6666	0.7500
		Neutral	0.6800	0.6800	0.6800
	Case 3	Positive	0.8135	0.9056	0.8571
		Negative	0.8888	0.8888	0.8888
		Neutral	0.7368	0.5600	0.6363
	Case 4	Positive	0.8524	0.9811	0.9122
		Negative	1.000	0.7777	0.8750
		Neutral	0.8421	0.6400	0.7272
Food	Case 1	Positive	0.8607	0.8717	0.8662
		Negative	0.7894	0.6250	0.6976
		Neutral	0.6842	0.7647	0.7222
	Case 2	Positive	0.8625	0.8846	0.8734
		Negative	0.9444	0.7083	0.8095
		Neutral	0.7105	0.7941	0.7499
	Case 3	Positive	0.8536	0.8974	0.8750
		Negative	0.8750	0.5833	0.6999
		Neutral	0.6578	0.7352	0.6944
	Case 4	Positive	0.8658	0.9102	0.8875
		Negative	0.9473	0.7500	0.8372
		Neutral	0.7428	0.7647	0.7536
Movie	Case 1	Positive	0.8686	0.8865	0.8775
		Negative	0.6904	0.6744	0.6823
		Neutral	0.6129	0.5937	0.6031
	Case 2	Positive	0.8613	0.8969	0.8787
		Negative	0.7045	0.7209	0.7126
		Neutral	0.6666	0.5625	0.6101
	Case 3	Positive	0.8571	0.8659	0.8615
		Negative	0.7142	0.6976	0.7058
		Neutral	0.5937	0.5937	0.5937
	Case 4	Positive	0.8787	0.8969	0.8877
		Negative	0.7619	0.7441	0.7529
		Neutral	0.7096	0.6875	0.6984

[4] Q.Mei, X. Ling, M.Wondra, H. Su, and C.X. Zhai, “Topic sentiment mixture: modeling facets and opinions in weblogs,” In Proceedings of the 16th International Conference onWorldWideWeb, pp.171-180, 2007.

[5] Z-Y. Ming, T-S. Chua, and G. Cong, Z. Ming, T, “Exploring domain-specific term weight in archived question search,” In Proceedings of the CIKM, 2010.

[6] B. Pang, L. Lee and S. Vaithyanathan, "Thumb up? Sentiment Classification Using Machine Learning Techniques," In Proceeding s of the EMNLP, pp.79-86, 2002.

[7] K. Dave, S. Lawrence and D. M. Pennock, Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In Proceedings of WWW 2003, pp.519-528. 2003.

[8] T. Mullen and N. Collier, Sentiment analysis using support vector machines with diverse information sources. In Proceedings of EMNLP 2004, pp.412-418, 2004.

[9] Charles E. Osgood, George J. Succi and Percy H.Tannenbaum, The Measurement of Meaning, University of Illinois. 1957.

[10] Changhua Yang, Kevin Hsin-Yih Lin, and Hsin-Hsi Chen. Emotion classification using web blog corpora. In WI ’07: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, pp.275–278, 2007.

[11] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In ICML ’01: Proceedings of the Eighteenth International Conference on Machine Learning, p. 282–289, 2001.

[12] Alec Go, Lei Huang, and Richa Bhayani, Twitter sentiment analysis. Final Projects from CS224N for Spring 2008/2009 at The Stanford Natural Language Processing Group, 2009.

[13] S.-H. Cho and H.-B. Kang, “Text Sentiment Classification for SNS-based Marketing Using Domain Sentiment Dictionary,” Proc. Int. Conf. on Consumer Electronics(ICCE), Las Vegas, 2012.

[14] L. Zhuang, F. Jing, and X.-Y. Zhu. “Movie review mining and summarization.”, In Proc. of CIKM ’06, pp. 43–50, New York, NY, USA, 2006. ACM.