Advanced Science and Technology Letters
Vol.95 (CIA 2015), pp.102-110
http://dx.doi.org/10.14257/astl.2015.95.18

A SNS Message Type Classification System Using Language Independent Features and Modified ECCD

Yang-Ha Chun[1,1],

[1] School of Computer Science & Engineering,
Yongin University, Gyeonggi 134, Korea
E-mail: yangha00@yongin.ac.kr

Abstract. This study proposes a SNS message type classification system combining language independent and dependent features that can be used in short message for type classification in social network service environments and verifies the effectiveness of this system. The language independent features are the Metadata of SNS message and the language dependent features are the bag-of-word selected by the modified ECCD feature selection method reflecting the short message characteristics. The experiment shows that the proposed system has better performance than comparable methods which use other language independent features or dependent features only.

Keywords: Twitter, SNS (Social Network System), Classification, Feature Selection, ECCD

1 Introduction

Most of the category-based information retrieval services provided by SNS are based on the category index which is directly assigned by the user who creates the message or system administrator who operates SNS portal service. However, the subjective judgment by the user or system administrator sometimes assigns messages to the wrong categories, which can cause error in applying the category index in information retrieval services. Thus, in order to provide a more efficient information retrieval services, an automatic message classification model according to the message type and context is required.

Existing term-based document classification researches, in general, classify documents according to selected features based on the bag-of-word approach [1]. However, this approach has a disadvantage in that it is language dependent and a POS-tagger and dictionary is required to extract term-based features [2]

Also, feature selection method is the term based that is selecting terms in document which is effective in classifying document in order to create the effective classification model. Common feature selection methods for terms are mutual

[1] Tel : +82-10-8984-7658
E-mail address: yangha00@yongin.ac.kr (Yang-Ha Chun)

ISSN: 2287-1233 ASTL

information (MI), information gain (IG), x^2 (Chi-square), ECCD (Entropy based Category Coverage Difference criterion), etc. [3][4][5].

Feature selection can be achieved by using correlation of inner category like mutual information or by using correlation of inner and inter categories such as IG, x^2 and ECCD. In particular, ECCD feature selection method of term base uses not only existence of terms that inner category but also Shannon entropy which considers the frequency of term presence in each document, which enables to evaluate terms more accurately[4][5].

However, in order to use feature selection method, there are problems in that the ratio of each category of documents that are used in selecting terms must be similar and there must exist at least one selected term in the document is predicted. Moreover, SNS messages are composed of only 140 characters [6].

On the other hand, using only independent feature which is Metadata in SNS message [7], it is hard to classify message type which often reflects strong pattern of users create message such as “message create time”, “length of message”, “frequency of use of special characters”, etcs. Also, message contents will not be reflected in the message type classification.

In order to solve the problems that we mentioned above, this study creates a model that integrates language independent features and dependent features. And proposes the SNS message type classification system that is possible to classify SNS message into four types (“News”, “Opinion”, “Event/Deal” and “Private”).

2 SNS Message Type Classification System

2.1 System Description

Fig.1 shows the system architecture of SNS message type classification system.

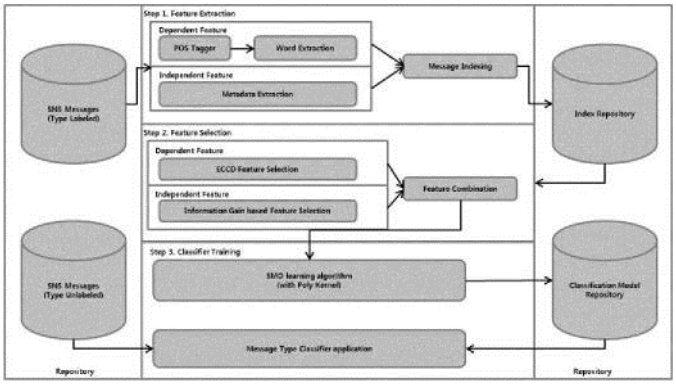

Fig. 1. System Architecture

Repository has "SNS Messages", "Index Repository" and "Classification Model Repository". SNS message are the row data of twitter that are crawled. These messages are split into "SNS Messages (Type Labeled)" that are labeled and "SNS Messages (Type Unlabeled)" that are not labeled.

In part of "Step 1. Feature Extraction", SNS message is collected then it processing incrementally. And "Step 1. Feature Extraction" is composed with extraction processing of dependent features and independent features.

Extraction of dependent feature processing is processed by "POS Tagger" and "Word Extraction", while extraction of independent features processing are processed by "Metadata Extraction".

Dependent feature extraction process of "Step 1. Feature Extraction" extracts term (a common noun, aproper noun) from word based on given condition determined from "POS Tagger" and "Word Extraction" from SNS message and delivers these extracted term to "Message Indexing". Then, "Message Indexing" saves indexing information of each message into "Index Repository".

Also, independent feature extraction process acquires refined information of Metadata which is collected from SNS message and refined by "Metadata Extraction".

The meaning of Incremental in "Step 1. Feature Extraction" is to update the information of "Index Repository" whenever SNS message is crawled.

"Step.2 Feature Selection" and "Step 3. Classifier Training are performed per certain classification update period which is set by administrator. "Step 2. Feature Selection" composed of "ECCD Feature Selection" and "Feature Combination", also "Step 3. Classifier Training" composed of "SMO learning algorithm" and "Message Type Classifier".

"Processing-2" can be divided into feature selection and learning classifier process. Firstly, feature selection process is comprised with dependent feature and independent feature selection. Dependent feature selection is to create meaningful set of term, "bag-of-word," using the technique of modified ECCD on "ECCD Feature Selection Module (for Dependent Feature)" which uses information of "Dependent Feature Information Repository" as input data. Independent Feature selection uses technique of information Gain on "Feature Selection Module (for Dependent Feature)" which uses information of "Independent Feature Information Repository" as input data.

Secondly, process of learning classifier is to combine selected dependent feature and independent feature in order to create final set of feature which will be used in the learning classification model. Inputting the final set of feature, "Learning Classifier Module" learns and creates SNS Message Type Classifier model.

'Step 3. Classifier Training' is composed of 'SMO learning algorithm' and 'Message Type Classifier application'.

2.2 Language Independent Features

The classification model proposed in this study uses only language independent features. The language independent features are divided into two parts: inner features existing in tweeted messages and outer features existing in the metadata information

of the tweeted messages. The feature extraction procedure uses Twitter API (twitter4j-2.2.5)[8] and regular expression of the JAVA program language. Details of the extracted features are shown in [Table 1].

Table 1. Details of the Extracted Independent Features

Feature	Type	Description	Data Type
f1	Outer	message writing time flag (Weekend)	binary (true/false)
f2	Outer	message writing time flag (Work Hour<08:00~18:00>)	binary (true/false)
f3	Outer	message writing time (Hour<00~23>) value	int (0~23)
f4	Outer	ReTweet message flag (RT Count>0 then true).	binary (true/false)
f5	Outer	ReTweet count of message	int (0<~)
f6	Inner	message length	Int (0~140)
f7	Inner	message length ratio = f6/(max message length)	double (0.0~1.0)
f8	Inner	space count	int (0~140)
f9	Inner	space count ratio = f8/(max message length)	double (0.0~1.0)
f10	Inner	Number of “RT”s in message	int (0~70)
f11	Inner	·Starts With “RT” in message″ flag.	binary (true/false)
f12	Inner	Number of “@”s in message	int (0~140)
f13	Inner	·Starts With “@” in message″ flag.	binary (true/false)
f14	Inner	URL in message flag.	binary (true/false)
f15	Inner	Number of URL″s in message.	int (0~25)
.....			
F42	Inner	Emoticon length ratio = (Emoticon length sum)/(message length)	double (0.0~140.0)

2.2.1 Outer Features

Outer features consist of features extracted from the metadata of SNS messages with Twitter API. The extracted features are the creation time of tweeted message and number of the message RTs. The creation time of the tweeted message is used in generating features f1~f3, whereas the number of message RTs is used in generating features f4~f5.

2.2.2 Inner Features

Inner feature extraction uses regular expressions and the matcher function of JAVA for the tweeted message. Regular expression extracts information, such as url, “@”, “RT”, parenthesis, phone number, emoticon and independent consonants or vowels.

2.3 Language Dependent Feature (bag-of-word)

2.3.1 Modified ECCD Feature Selection Method.

The technique of modified ECCD feature selection [9], selecting the term con-sidering both ratio of SNS message type and ECCD feature selection method, and

therefore, reduce the problem when there is a difference among ratio of each message type of SNS.

"Modified ECCD" feature selection method follows preceding ECCD feature selection method and, in addition, select appropriate number of term considering the ratio of message type. This method finds the value of ECCD(t_j,c_k) of each type selection word considering Shannon entropy E(t_j) as following 1.

(1)

Σ ( ) Σ (2)

( ) – (3)

( ) (4)

Final creation of set of selection term is such that the value of ECCD(t_j,c_k) is listed in descending order for each type and appropriate number of term is decided. Then, last step for final creation of set is to unionize the term that are chosen for each type. For appropriate number of term can be found as in (7).

: Message Type percentage of the total Message (5)

: selection word Number of settings (6)

(7)

2.3.2 Language Dependent Feature Value

A set of term selected by language dependent features are used in learning classification model and the value of term is frequency of presence of term in message.

3 Experiments

3.1 Data

Tweet messages in Korean from 2012-10-14 to 2012-12-08 were collected for the experiment. Among the collected tweet messages, 1600 messages were extracted which are labeled as one of four types (News, Opinion, Event/Deal and Private).

3.2 Method

This experiment uses a classification algorithm provided by Weka 3.7.7 [6]. Experimental evaluation was conducted using 10-fold cross-validation with precision, recall and F-measure. From the experiment, among the various classification algorithms, such as Logistics, Multilayer-Perceptron, NaiveBayes, J48 and SMO, we found that SMO algorithm showed the best performance in classification accuracy. Therefore, in experimental environment, we use SMO algorithm.

3.3 Results

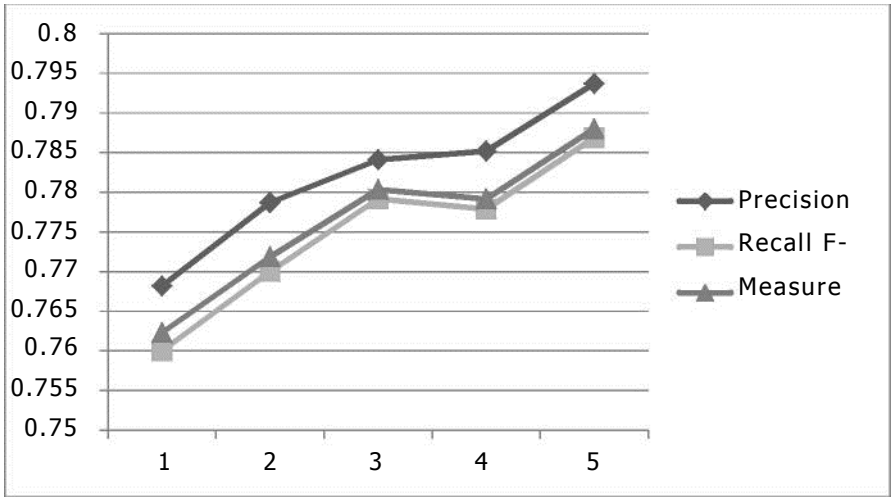

Fig. 2. Experimental Results Using Independent Features

Table 2. Details of the Extracted Independent Features Accuracy

=== Detailed Accuracy By Class ===				
	Precision	Recall	F-Measure	Class
	0.658	0.79	0.718	news
	0.771	0.708	0.738	opinion

	0.758	0.69	0.723	deal/event
	0.987	**0.96**	**0.973**	private message
Weighted Avg.	0.794	0.787	0.788	

According to Fig. 2, using language independent feature model of the SMO algorithm shows the accuracy up to 79% for the 1600 messages dataset (each message category has 400 dataset). This is better performance than that of comparable research [3][6] which uses only term-based features and results in an accuracy of 65% ~ 75%.

As shown in Table. 2, comparatively "private message" shows higher accuracy than the others.

Table 3. Details of the Extracted Dependent Features Accuracy

=== Detailed Accuracy By Class ===				
	Precision	Recall	F-Measure	Class
	0.652	0.543	0.592	news
	0.534	0.41	0.464	opinion
	0.859	0.668	0.751	deal/event
	0.522	0.848	0.646	private message
Weighted Avg.	**0.642**	**0.617**	**0.613**	

In this case, the accuracy is lower compared to the study result of web news and, as shown in table 3, prediction accuracy is lower for message of "opinion" and "news" type.

Table 4. Details of the Extracted Combine Independent and Dependent Features Accuracy

=== Detailed Accuracy By Class ===				
	Precision	Recall	F-Measure	Class
	0.708	0.775	0.74	news
	0.758	0.75	0.754	opinion
	0.836	0.793	0.814	deal/event
	0.997	0.965	0.981	private_message
Weighted Avg.	0.825	0.821	0.822	

It has better accuracy than that of comparable methods which uses only language independent features or dependent features model. Moreover, as shown in Table 4, according to the message type, standard deviation of classification accuracy is small then other independent or dependent features model.

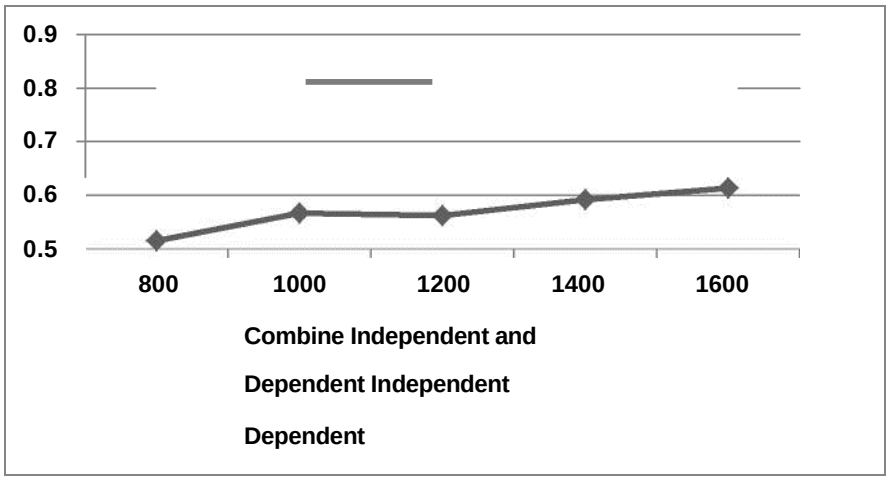

Fig. 3. Compare F-Measure

The model uses dependent and independent features together, and these two different features complement each other in classification process and improve the accuracy.

4 Conclusions and Future Work

In this paper, we proposed a classification method using independent and dependent features that are effective in SNS message type classification. The evaluation result of 4 type message classification shows improved performance compared to the language independent and dependent based classification model. The reason for improved accuracy is that it would be difficult for term-based approach to extract meaningful features from short SNS. Furthermore, integrating language independent feature with language dependent features created by "Modified ECCD Feature Selection method" in one model is more effective than using language independent features only.

However, comparing the performance with the previous methods with the same data directly other than Korea language may be needed for the future work

References

1. Salton, G. "Automatic processing of foreign language documents," Journal of the American Society for Information Science, 21(3), pp. 187-194 (1970)
2. Hongguang Zheng, Nobuhiro Kaji, Naoki Yoshinaga, and Masashi Toyoda, A Study on Microblog Classification Based on Information Publicness, In Proceedings of the 4th Forum on Data Engineering and Information Management. Kobe, Hyogo. (2011)
3. Sanasam Ranbir Singh, Hema A. Murthy, Timothy A. Gonsalves, "Feature Selection for Text Classification Based on Gini Coefficient of Inequality" JMLR:Workshop and Conference Proceedings, pp. 76-85, (2010)

4. Christine Largeron, Christophe Moulin, Mathias Gery, “Entropy based feature selection for text categorization” Proceedings of the 2011 ACM Symposium on Applied Computing, pp. 924-928, (2011)
5. C. E. Shannon, “A mathematical theory of communication” ACM SIGMOBILE Mobile Computing and Communications Review, Volume 5 Issue 1 (2001)
6. Aixin Sun, “Short Text Classification Using Very Few Words”, Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval , pp. 1145-1146 (2012)
7. ChungSeok Han, Sangyong Park, Soowon Lee, "SNS Message Type Classification Using Language Independent Features", Sixth International Conference on Information (2013)
8. Twitter API, https://dev.twitter.com/
9. Chungseok Han, Sangyong Park, Soowon Lee, “A Document Classification System Using Modified ECCD and Category Weight for each Document”, information processing society journal B Volume 19 Issue 4, pp.237-242 (2012.08)

