IJCLA VOL. 2, NO. 1–2, JAN-DEC 2011, PP. 285–300
RECEIVED 02/11/10 ACCEPTED 21/12/10 FINAL 11/02/11

Features of Latinate Etymologies on the Tasks of Text Categorization

WANYIN LI AND ALEX CHENGYU FANG

City University of Hong Kong, Hong Kong

ABSTRACT

In European languages development, new words and/or new terms are mostly formed using Latin and/or Greek word elements. Linguists have proved the importance of these elements in forming an important part of LSP word stock in different subject fields such as computer science and medicine. This article attempts to show the effectiveness of using words with Latinate etymologies as features for the tasks of text categorization (TC). One essential step for training classifiers in TC to be more accurate is the effective feature selection. Lots of methods for feature selection have been discussed by computer scientists on the basis of information theory and from the perspective of statistics. To cope with the bottleneck of high dimensionality from bags of words (BOW), the core features to be discussed in this article are selected according to etymological information of words and therefore are drastically different from existing feature selection methodologies. The result is analyzed from evaluation schemes including accuracy, precision, recall, and F-measure. The experiment shows that the Latinate words as discriminative features are effective to reduce the dimensionality of the feature space and outperform other feature combinations in F-value of up to 98% when using one of the state-of-the-art methods, i.e., Support Vector Machine from WEKA.

KEY WORDS: *Text Categorization, Feature Selection, Etymology, Latinate Token, Lexicon, Classifier, BOW, Naïve Bayes, Decision Tree, SVM.*

1 INTRODUCTION

With the explosive growth of the Internet, more research efforts are required for the task of text categorization to improve accuracy and efficiency. Two key steps involved in text categorization include the selection of features and the training of classifiers. Reports on feature selection methods with the purpose of scaling down the feature space are usually based on statistical theory and machine learning such as information gain [7, 8, 12, 25], mutual information [13, 25], Chi-square [8, 12, 25] etc. It has been stated [21] that most of the dimensions from bags of words (BOW) are not typically relevant to text categorization and even introduce noise features even though they are said to be statistically important, hence eventually hurt the performance of the training classifiers. Actually, even the proved most successful training classifier for the tasks of text categorization like Support Vector Machine (SVM) is inefficient when directly taking the words/phrases based on the statistical theory as features [27]. It is thus necessary to select distinguishing features according to a methodology different from the conventional statistically based machine learning methods.

This article, taking a different direction from the previous research studies, proposes the selection of core features according to words of a Latin origin based on a lexicon that contains etymological information. Six feature sets are set up including BOW, Latinate words, and the content words from nouns, verbs, adjectives and adverbs. Three selected classifiers, namely SVM, Bayes, and decision tree, are then applied on the above feature sets to find their supportive votes to each of the feature sets. The experiments consistently indicate that the feature set of Latinate words outperforms the other feature sets based on the three selected classifiers.

The rest of this article is organized as follows. Section 2 investigates the rationale behind the use of Latinate words as discriminative features. Section 3 reviews the related studies. Section 4 describes the approach on BNC written texts categorization. Section 5 discusses the result of the experiment and Section 6 concludes the article.

2 WHY LATINATE WORDS

Two lexicographical resources are used in this study. One is a 100 million word collection named British National Corpus (BNC) with

90% of its content coming from written texts. The written part comes from regional and national newspapers, specialist periodicals and journals, academic books and popular fiction, published and unpublished letters and memoranda, school and university essays. The version used in this paper contains texts primarily in 1990's. Another resource contains total 249,331 entries, named Collins English Dictionary (CED), which is a collection of total 128 different languages of etymological knowledge including Latin, French, and Greek etc for contemporary English. The etymological resources in CED are identified by three different tags which output total of 48,593 entries as the final etymological origins.

Recall in the tasks of TC, features drawn from top ranked list based on statistical feature selection face problem that the selection algorithms are either over-relied or mislead by infrequent terms. The way is to select a small number of such feature straightly which may be sufficiently efficient and preserves the relevant information to the texts without utilizing statistics based feature selection. Similar idea is supported by [9] Forman that "additional complexity of feature selection can be omitted for many researchers who are not interested in feature selection, but simply need a fixed and easily replicable input representation".

In summary, the intuition behind the use of Latinate words as features is based on three observations. The first observation is that the linguists have proved that over 70% of the words used in modern English have been borrowing extensively from other languages, especially from Latin, French and Greek [10, 11, 23]. The second observation is that certain contexts from special domains (such as the domain of medicine) would be more likely to turn to certain type of foreign words (such as Latinate words) [4, 24]. The third observation based on a previous work [5] proves that even the density of Latinate words from the distinct text domains performs acceptable to distinguish the spoken and written BNC texts in an average precision rate of 80%. This work aims to further disclose whether the words having Latinate etymologies are effective in distinguishing BNC written texts.

3 PREVIOUS STUDIES

Feature selection is the first key step in a successful task of text categorization. The previous works for this step were mainly based on machine learning theory to pick up the features which are significant in

statistical calculation. When reviewing previous studies with respect to this step, we have in mind two motivations: (1) what indexing terms were selected as features and (2) what weighting schemes were applied to score the selected terms. The features have been typically discussed in relevant studies including single tokens/words [12], keywords [1, 25], bi-grams/n-grams [17, 20], noun phrases [3, 26], and syntactically bounded phrases [16, 18] with respect to the predefined/learned syntactic patterns. Among these features, it has been reported that "stemmed or un-stemmed single words as features give better classifier performance compared with other types of features" [3]. With the summarization on these reviewing facts, this paper makes the typical choice of the indexing terms as different individual tokens like BOW, Latinate tokens, and stemmed tokens from four types of content words.

For weighting schemes, studies [8, 22] have reported that traditional feature selection methods may be over relied when identifying statistical significant features. Likewise, Olsson [12] reported that χ^2 are known to be misled by infrequent terms. Liao [3] also concluded that LOG(tf).IDF as feature weight gives better classifier performance than other types of feature weighting schemes. In this work, we utilize the interface of StringtoWordVector provided by WEKA to transform the selected features into tf*idf vectors as the input of learning algorithms supported by WEKA as well.

To the end of learning algorithms on the tasks of TC, the theory from machine learning is usually applied, like Naïve Bayes Classification [6, 25], Support Vector Machine [6, 8, 15, 25], KNN [6, 8, 12], Decision rule/tree [2, 19] etc. Damerau and Weiss stated in [2] that machine generated decision rules able to compete with human performance in text categorization, whereas Joachims proved that better result can be achieved by using Support Vector Machines [14]. J. Wulandini [6] also showed SVM performs the best compared with Naive Bayes and KNN with 92.5% of accuracy. In this work, we select Naive Bayes, C4.5 Decision Tree, and SVM from WEKA as the learning algorithms to evaluate the candidate feature sets.

4 THE PROPOSED METHOD

The architecture of machine learning based text categorization consists of two main parts (1) Feature selection procedure to identify candidate features; then reduce the feature space using filters; and finally transform the selected textual features into numerical values. (2) The

learning procedure to learn a training model and classify the testing data. Fig. 1 gives the full picture to describe the standard text categorization procedure. This work focuses on the first part, hence, we describe the procedures 1.1, 1.2 and 1.3 which while fulfilled will append the candidate features into the scalable features database discussed in the next section.

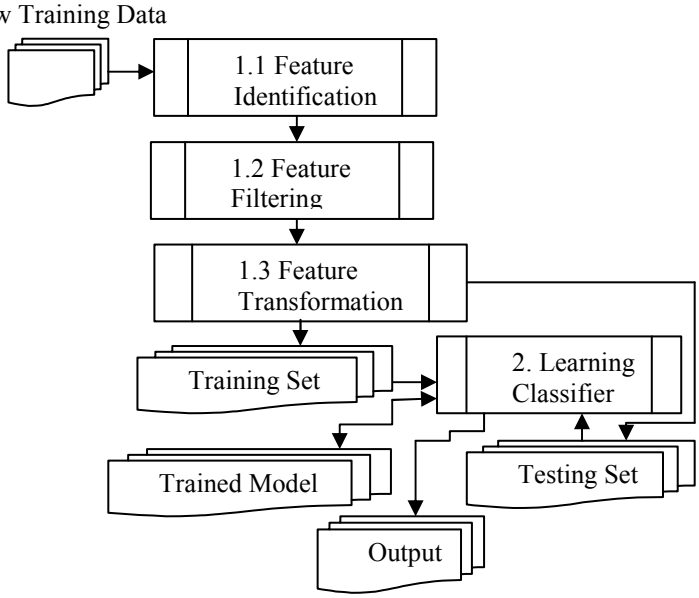

Fig. 1. Architecture of ML-based TC system

4.1 BUILDING UP THE FEATURES DATABASE

In order to compare with the other types of features, we include BOW, Noun/Verb/Adjective/Adverb tokens as well. Fig. 2 is the framework to identify the different types of candidate features mentioned in the process 1.1 in Fig. 1.

BNC has been well annotated in terms of tokenization, lemmatization as well as part-of-speech (POS) tags. Upon the feature identification procedure, in process 2.1, all headwords (HWs) with the tag of POS are directly extracted. As BNC corpus has been annotated

by stemmed tokens for each of the HWs, process 2.2 employs a top-2000 frequent wordlists which is calculated based on the whole corpus to filter out the most frequent tokens. The filtered HWs are fed with POS tags to extract the features of Noun/Verb/Adjective/Adverb tokens in the process 2.5.

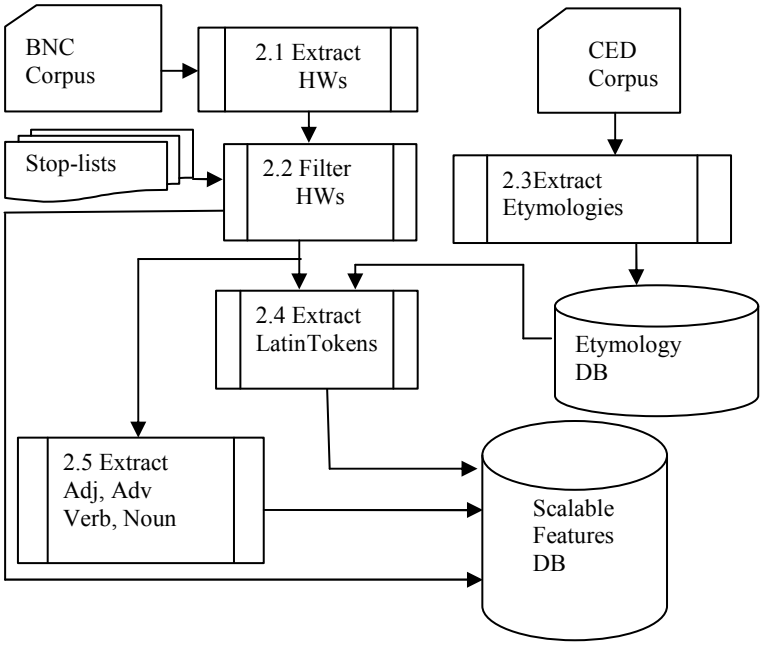

Fig. 2. Framework to build up the features DB

Coming back to the extraction of Latinate tokens, in process 2.3, CED dictionary is firstly utilized to look for root words with etymological origins. These three types are exemplified as follows:

1. cabezon ety Spanish, from cabeza head, ultimately from Latin caput
2. agglutinin ety C19: #7 agglutinate
3. cabinet-making sub head cabinet-maker

The first type assigns Latin as the etymology of the word “cabezon”. In the second case, the etymology of “agglutinin” is assigned by looking for the etymology of “agglutinate”. The third type takes the etymology

of "cabinet-maker" as the etymology of "cabinet-making". All of the three types of tags when traced extract a total of 48,593 etymological entries. For this writing, all entries having Latinate etymology will be further selected from the etymology database as the input to extract Latinate tokens in the process 2.4. The output from the processes 2.2, 2.4, and 2.5 will be appended to the feature database as the input for the selected classifiers in the next section.

For the process 1.2 of feature filtering in Fig. 1, we employ the default filter from WEKA to remove the functional words as well as the words with the frequencies less than 3 for the BOW feature sets to reduce the features space.

For the feature transformation in the process 1.3 of Fig. 1, the standard tf*idf supported by WEKA is applied to index the terms.

4.2 Learning Classifier

To show tokens with Latinate etymologies the discriminative features, Naïve Bayes, C4.5 Decision Tree (J48 in WEKA), and SVM (SOM in WEKA) are selected as learning classifiers to evaluate the selected features. All these classification algorithms are trained using WEKA's default parameter setting such as in SMO using polynomial kernel function, the complexity parameter as C = 1.0 and exponent = 1.0.

5 Experimental Design

The experiment is designed with two testing approaches applied. Firstly, the training models with respect to each learning classifiers are built up based on the train-and-test approach. With the aim of keeping the optimization possible, in the train-and-test approach, each category in the initial BNC corpus has been randomly split into 80% of each as training set to build up the classifier, and an exclusive 20% of each as testing set to test the effectiveness of the classifier. The above splitting is repeated in 5 times. Alternatively, the approach of 5-fold cross validation supported by WEKA is also employed to obtain a macro-averaged performance over the different classes. Table 1 shows the number of even-sized texts under each of the written categories with one sample train-test splitting.

Table 1. Number of texts in written categories.

	Fiction	News	Otherpub	Unpub
Train	2,986	2,856	2,946	2,911
Test	811	826	831	761

5.1 FEATURE SETS

Six datasets such as BOW tokens, Latinate tokens, adjective tokens, noun tokens, verb tokens, and adverb tokens are used in the experiment. Table 2 shows the number of instances in different datasets used as input for the three selected classifiers as well as the time cost by the classifiers to build up the training models.

As stated above, there are total 48,593 etymological entries extracted from CED. Taking the entry list as an input to extract the Latinate tokens, total 2,338 instances of Latinate tokens are returned from the four categories. In the experiment, the number of features in the set of Latinate tokens are approximately the order of 10^5, in the other sets equal to the number of instances times 50. We follow the complexity analysis scheme raised in [21] by using a percentage of features when evaluating the performance of the different types of features, because of the absolute size of tokens vary greatly in the above different feature sets.

Table 2. Number of instances in each dataset.

	No. of Instances		Time to Train Model (minutes)		
	Train	Test	Bayes	J48	SMO
Latinate Tokens	2,338	645	.073	0002	0.17
BOW	75,698	15,140	0006	0469	1982
Adj. Tokens	20,873	5,651	1.13	0093	0099
Noun Tokens	51,409	13,603	3.08	0260	0446
Verb Tokens	34,858	9,291	2.10	0178	0286
Adv. Tokens	12,070	3,363	0.63	19.5	0026

5.2 EVALUATION

We wish to compare the impact of different feature sets based on different evaluation scheme such as precision, recall, Break-even-point (BEP), and macro-average F-measure. With respect to the contingency table described in Table 3, the above named evaluation scheme is defined in formula (1)-(4).

Table 3. Contingency table.

		System Yes	Prediction No
Standard Answer	Yes	*TP*	*FN*
	No	*FP*	*TN*

$$precision: P = \frac{TP}{TP + FP} \quad (1)$$

$$recall: R = \frac{TP}{TP + FN} \quad (2)$$

$$BEP: BEP = \frac{P + R}{2} \quad (3)$$

$$F - measure: F = \frac{2PR}{P + R} \quad (4)$$

Macro-average F-measure is the average on F scores of all the classes.

TP: True Positive results; *FN*: False Negative results
FP: False Positive results; *TN*: True Negative results

5.3 EXPERIMENTAL RESULTS

Table 4 shows the Macro-averaged precision (P), recall (R), BEP (B), and F-value (F) over four categories against the above six features with 5-fold cross-validation for the selected classifiers Naïve Bayes (NB), J48, and SMO respectively.

Table 4. Average performance over four categories with respect to the feature sets.

NB	BOW	Latin	Adj.	Noun	Verb	Adv.
P	.654	**.845**	.652	.608	.624	.708
R	.661	**.814**	.671	.633	.624	.686
B	.657	**.830**	.662	.621	.624	.697
F	.657	**.820**	.662	.620	.624	.697
J48	BOW	Latin	Adj.	Noun	Verb	Adv.
P	.693	**.899**	.525	.540	.542	.566
R	.680	**.890**	.530	.542	.552	.575
B	.687	**.895**	.528	.541	.541	.571
F	.686	**.890**	.524	.540	.545	.569
SMO	BOW	Latin	Adj.	Noun	Verb	Adv.
P	.709	**.983**	.675	.625	.670	.636
R	.690	**.980**	.684	.636	.669	.649
B	.700	**.982**	.680	.631	.670	.643
F	.700	**.982**	.679	.630	.670	.642

Fig. 3 and Fig. 4 show the performance for the conducted feature sets with respect to the three selected classifier in macro-precision and macro-F-value.

The consistent agreement is achieved from both Fig. 3 and Fig. 4 that the best performance was achieved with the feature set based on Latinate tokens. The same conclusion also holds true for all the three selected learning algorithms. Especially in SMO, both the precision and macro-F-value vary in the range of 88% to 98% with the increase of the number of features, which outperform other features such as noun tokens which produced the precision of 62.5% and F-value of 63% on average. The results also disclose that the performance with Latinate tokens as features increases when the number of features in the feature set increases, while this characteristic is not significant for the other feature sets used in our experiments. Except for Latinate tokens, no consistent conclusion can be drawn for the other feature sets with respect to the three algorithms. SMO and J48 give better performance for BOW compared with the features of noun/verb/adjective/adverb tokens, while the worst macro-performance was recorded for adjective tokens.

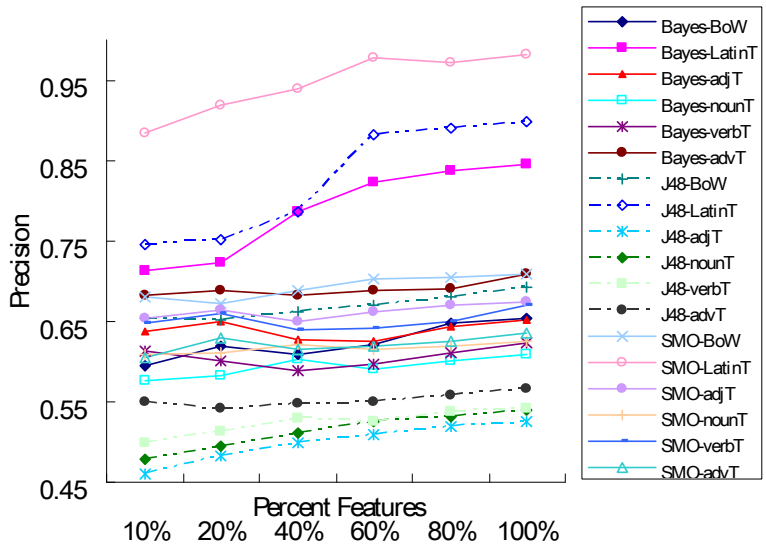

Fig. 3. Precision on six feature sets

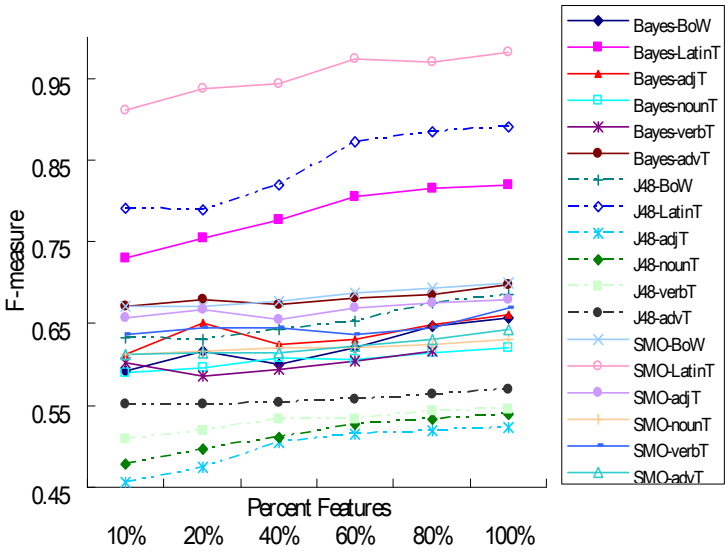

Fig. 4. F-value on six feature sets

6 CONCLUSION AND FURTHER DISCUSSION

This work has tested a number of different feature sets using the learning algorithms of Naïve Bayes, Decision Tree, and SMO from WEKA. The experiments reported in this article concludes that features based on Latinate tokens are superior to the features of BOW as well as stemmed noun/adjective/adverb/verb tokens, in both accuracy and training efficiency. This study finds almost the best performance in the term of F-measure over 98% with the features of the stemmed Latinate tokens reported so far. The result additionally confirms the previous findings [6, 14] that SVM classifier achieves an outstanding performance on the tasks of TC.

Table 4 in Section 5.3 shows that the Latinate tokens as features outperform other selected feature sets. When we look back to try to seek any explanation for the above special result, two reasons appear to be noteworthy for further study. The first reason is that the number of instances is inconsistent in the named feature sets. The number of instances is in terms of thousands in the feature set of Latinate tokens, while it is in terms of tens of thousands in the other four feature sets (Adj/Noun/Verb/Adv). For this observation, we make the number of instance in each of the above four feature sets evenly by randomly selecting thousands from each of them. The second reason is that the number of categories tested in the experiments is small (four in total). To this end, we extend the experiments into eight categories including ACPROSE (academic prose), CONVRSN (conversation), FICTION (fiction), NEWS (news), OTHERPUB (other publication), NONAC (non-academic propose), OTHERSP (other speech), UNPUB (unpublished writing). The results shown in Table 5 agree with the results from Table 4 and in both cases Latinate tokens perform the best among the five feature sets. This thus reinforces our conclusion that Latinate tokens constitute an effective discriminative feature set for the tasks of text categorization.

ACKNOWLEDGEMENT. The research reported in this article was supported in part by research grants from City University of Hong Kong (Project Nos 7008002, 7008062 and 9610126).

Table 5. Performance over eight categories with even number of instances based on the six feature sets.

NB		Convrsn	Othersp	Acprose	Nonac	Fiction	News	Otherpub	Unpub
Latin	P	.923	.734	.76	.779	.799	.853	.853	.693
	R	.712	.715	.73	.684	.667	.735	.736	.72
	B	.817	.725	.745	.732	.733	.794	.795	.707
	F	.804	.724	.745	.728	.727	.789	.790	.707
Adj	P	.702	.671	.182	.194	.518	.573	.314	.615
	R	.745	.606	.522	.398	.714	.592	.473	.521
	B	.723	.638	.352	.296	.616	.582	.393	.568
	F	.723	.637	.269	.261	.600	.582	.377	.564
Noun	P	.694	.696	.412	.559	.709	.692	.594	.493
	R	.786	.581	.648	.330	.878	.555	.309	.244
	B	.74	.567	.469	.511	.703	.588	.526	.540
	F	.737	.538	.462	.507	.703	.569	.517	.536
Verb	P	.601	.508	.361	.391	.612	.528	.467	.325
	R	.794	.598	.439	.414	.790	.586	.494	.580
	B	.697	.553	.400	.402	.701	.557	.481	.452
	F	.684	.549	.396	.402	.689	.555	.480	.416
Adv	P	.727	.631	.548	.265	.651	.641	.400	.282
	R	.778	.586	.598	.368	.789	.567	.375	.333
	B	.752	.608	.573	.316	.720	.604	.387	.307
	F	.752	.608	.572	.308	.713	.602	.387	.305
J48		Convrsn	Othersp	Acprose	Nonac	Fiction	News	Otherpub	Unpub
Latin	P	.589	.690	.736	.821	.736	.821	.916	.826
	R	.621	.723	.488	.738	.932	.833	.928	.870
	B	.605	.706	.612	.779	.834	.827	.922	.848
	F	.604	.706	.587	.777	.822	.827	.921	.847
Adj	P	.419	.314	.336	.304	.407	.412	.261	.321
	R	.402	.334	.356	.305	.452	.373	.250	.312
	B	.411	.324	.346	.304	.430	.392	.255	.316
	F	.410	.323	.346	.304	.428	.391	.255	.316
Noun	P	.498	.275	.386	.221	.237	.229	.142	.217
	R	.394	.427	.339	.154	.142	.435	.141	.130
	B	.446	.351	.362	.187	.189	.332	.141	.173
	F	.439	.334	.361	.181	.177	.300	.141	.162
Verb	P	.408	.337	.317	.265	.372	.281	.239	.315
	R	.387	.337	.364	.235	.416	.361	.218	.278
	B	.397	.337	.341	.250	.394	.321	.228	.297
	F	.397	.337	.338	.249	.393	.316	.228	.295
Adv	P	.505	.402	.354	.203	.330	.502	.219	.191
	R	.537	.502	.253	.170	.388	.526	.202	.195
	B	.521	.452	.303	.186	.359	.514	.211	.193
	F	.521	.446	.295	.185	.356	.514	.210	.193

SMO		Convrsn	Othersp	Acprose	Nonac	Fiction	News	Otherpub	Unpub
Latin	P	.800	.894	.722	.731	.880	.960	.933	.927
	R	.582	.766	.570	.633	.961	.873	.956	.653
	B	.691	.830	.646	.682	.921	.916	.944	.940
	F	.674	.825	.637	.678	.919	.914	.944	.939
Adj	P	.516	.418	.298	.531	.433	.344	.228	.264
	R	.673	.582	.257	.243	.720	.560	.339	.430
	B	.594	.500	.277	.387	.576	.452	.283	.347
	F	.584	.487	.275	.333	.541	.426	.272	.327
Noun	P	.617	.357	.446	.273	.622	.351	.280	.441
	R	.696	.596	.481	.237	.551	.591	.441	.375
	B	.656	.476	.463	.255	.586	.471	.361	.408
	F	.654	.446	.462	.253	.584	.440	.342	.405
Verb	P	.532	.400	.302	.314	.558	.323	.304	.298
	R	.698	.588	.305	.282	.607	.506	.392	.354
	B	.615	.494	.303	.298	.582	.414	.348	.326
	F	.604	.476	.303	.297	.581	.394	.342	.323
Adv	P	.601	.532	.502	.220	.476	.452	.264	.211
	R	.740	.696	.402	.283	.629	.632	.303	.264
	B	.671	.614	.452	.252	.552	.542	.283	.238
	F	.663	.603	.446	.247	.542	.527	.282	.234

REFERENCES

1. Anette, H.: A study on automatically extracted keywords in text categorization. In: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL, pp. 537-544 (2006)
2. Apte, C., Damerau, F. & Weiss, S.M.: Automated Learning of Decision Rules for Text Categorization. ACM Transactions on Information Systems, v.12, pp. 233-251 (1994)
3. Liao, C., Alpha, S. & Dixon, P.: Feature Preparation in Text Categorization. In: Proceedings of Australian Workshop of Data Mining in CEC (2003)
4. Cabré, T.M.: Terminology: Theory, Methods, and Applications. Amsterdam; Philadelphia: John Benjamins Pub. Co (1999)
5. Fang, A.C., Li, W.Y. & Ide, N.: Latin Etymologies as Features on BNC Text Categorization. In: Proceedings of 23rd Pacific Asia Conference on Language, Information and Computation, Hong Kong, China (2009)
6. Wulandini, F. & Nugroho, A.S.: Text Classification Using Support Vector Machine for Webmining based Spatio Temporal Analysis of the Spread of Tropical Diseases. In: Proceedings of International Conference on Rural Information & Communication Technology, Bandung Institute of Technology, Indonesia, pp. 189-192, (2009)

7. Lee C. & Geunbae, G.: Information gain and divergence-based feature selection for machine learning-based text categorization. Information Processing and Management: An International Journal, v.42 n.1, pp. 155-165 (2006)
8. Shang, W., Huang, H., Zhu, H., Lin, Y., Qu, Y. & Wang, Z.: A novel feature selection algorithm for text categorization. Expert Systems with Applications: An International Journal, v.33 n.1, pp. 1-5 (2007)
9. Forman, G.: Feature Selection for Text Classification. Chapman and Hall/CRC Press (2007)
10. Greenbaum, S.: The Oxford English Grammar. Oxford University Press (1996)
11. Hughes, G.: A History of English Words. Blackwell Publishers (2000)
12. Olsson, J.O.S. & Oard, D.W.: Combining feature selectors for text classification. In: Proceedings of the 15th ACM international conference on Information and knowledge management, November 06-11, Arlington, Virginia, USA (2006)
13. Pei, Z., Shi, X., Marchese, M. & Liang, Y.: Text Categorization Method Based on Improved Mutual Information and Characteristic Weights Evaluation Algorithms. In: Proceedings of the Fourth International Conference on Fuzzy Systems and Knowledge Discovery, Vol.4 (2007)
14. Joachims, T.: A Statistical Learning Model of Text Classification for Support Vector Machines. In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press, pp. 128-136 (2001)
15. Joachims, T.: Training linear SVMs in linear time. In: Proc. of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 217–226 (2006)
16. Furnkranz, J., Mitchell, T. & Riloff, E.: A case study in using linguistics phrase in text categorization on the WWW. In: Proceedings of AAAI/ICML Workshop (1998)
17. Kanaris, I. & Stamatatos, E.: Webpage genre identification using variable-length character n-grams. In: Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence, Vol.2, Washington, DC, USA, IEEE Computer Society (2007)
18. Lewis, D.D.: Feature selection and feature extraction for text categorization. In: Proceedings of Speech and Natural Language Workshop (1992)
19. Lewis, D.D. & Ringuette, M.: A Comparison of Two Learning Algorithms for Text Categorization. In: Third Annual Symposium on Document Analysis and Information Retrieval, pp. 81-93 (1994)
20. Mansur, M., UzZaman, N., Khan, M.: Analysis of n-gram based text categorization for Bangla in a newspaper corpus. In: Proceedings of the 9th International Conference on Computer and Information Technology, Dhaka, Bangladesh (2006)

21. Rogati, M. & Yang, Y.: High- performing feature selection for text classification. In: Proceedings of CIKM '02, pp. 659–661. ACM Press (2002)
22. Mukras, R., Wiratunga, N., Lothian, R., Chakraborti, S. & Harper, D.: Information Gain Feature Selection for Ordinal Text Classification using Probability Re-distribution. In: Proceedings of the Textlink workshop at IJCAI-07 (2007)
23. Roberts, A. H.: A Statistical Linguistic Analysis of American English. The Hague: Mouton (1965)
24. Stockwell, R. and Minkova, D.: English Words: History and Structure. Cambridge University Press (2001)
25. Wang, G., Lochovsky, F.H. and Yang, Q.: Feature selection with conditional mutual information maximin in text categorization. In: Proceedings of the 13th ACM International Conference on Information and Knowledge Management, ACM, Washington, DC, USA. pp. 342-349 (2004)
26. Zhang, W., Liu, S., Yu, C., Sun, C., Liu, F. & Meng, W.: Recognition and classification of noun phrases in queries for effective retrieval. In: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, USA (2006)
27. Zhang D. & Lee W.S.: Extracting key-substring-group features for text classification. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM Press, pp. 474–483 (2006)

WANYIN LI
DEPARTMENT OF CHINESE, TRANSLATION AND LINGUISTICS,
CITY UNIVERSITY OF HONG KONG,
HONG KONG
E-MAIL: <CLAIRELI@CITYU.EDU.HK >

ALEX CHENGYU FANG
DEPARTMENT OF CHINESE, TRANSLATION AND LINGUISTICS,
CITY UNIVERSITY OF HONG KONG,
HONG KONG
E-MAIL: <ACFANG @CITYU.EDU.HK >