F-term classification Experiments at NTCIR-6 for Justsytems

Masaki RIKITOKU
Justsystems Corporation Aoyama Bldg.
1-2-3 Kita-Aoyama, Minato-ku, Tokyo 107-8640, Japan
masaki_rikitoku@justsystem.co.jp

Abstract

We conducted the classification subtask at NTCIR-6 Patent Retrieval Task using a system based on three document classifiers, namely, a one-vs-rest SVM classifier, multi-topic classifier, and binary Naive Bayes classifier.

The multi-topic classifier was constructed on the basis of the maximum margin principle and applied to multiple F-term classification. From the experimental results, this multi-topic classifier yielded a higher F1 value than the one-vs-rest SVM in many cases.

In addition, we employed the one-vs-rest SVM classifier. The SVM classifier has certain drawbacks such as low recall performance and large learning time. In order to solve these problems, we used heuristics for achieving random reduction of a part of the negative examples and division of learning. These procedures lead to a reduction in learning time and improve the classification performance when appropriate parameters are set.

Keywords: *Machine Learning, Text Classification, Naive Bayes, SVM, Multi-topic Classification, Structured Classification*

1 Introduction

The classification subtask at NTCIR-6 Patent Retrieval Task is the task that multiple F-terms are automatically detected for a given patent document. A user can search a patent document by using this F-term index in a multifaceted manner [1].

This task has certain characteristics from the viewpoint of document classification. The F-term classification system has a tree structure. Each F-term is linked by tree edges. The upper level F-terms have broader concepts, while the lower level F-terms have more detailed concepts.

Furthermore, the number of multiple F-terms to be given is relatively large, and each F-term has a relation that is based on the classification system. we can consider that this task to be a multi-topic document classification task having structured outputs.

From this viewpoint, we apply three classification methods to this task, namely, the one-vs-rest SVM classifier (SVM), multi-topic classifier (MTC), and one-vs-rest Naive Bayes classifier (NB).

The MTC implementation for the multi-topic classification algorithm is based on the maximum margin principle [6],[9]. Unlike the other algorithms based on binary classifiers, we can expect this algorithm to produce an appropriate score for multiple F-terms and improve the low recall property of the SVM by using the one-vs-rest formulation.

This time, the number of tasks and training data are significantly larger than those at NTCIR-5. Thus, the execution of experiments is difficult if they are performed in a naive manner. For this reason, we apply certain heuristics for the one-vs-rest SVM. one of these is the reduction of a part of the negative training examples, while the other is division of learning.

This paper is organized as follows. In section 2, we describe the feature extraction of patent documents. In section 3, we provide a detailed description of the SVM, MTC, NB, and some heuristics for the SVM. In section 4, we describe the results of classification and analysis. Finally, in section 5, we provide the conclusion.

2 Feature selection

A patent document has structured data, applicant information, abstracts, claims, and so on. Therefore, it is possible to improve the classification performance by leveraging this structured data [4]. Meanwhile, the structured data yields high dimensionality in the feature space of the classifier. This will decrease the classification performance if the training data are relatively small.

In this task, the sizes of the F-term set and multiple F-terms to be provided are large, while the training data are relatively small in most cases. Therefore, we use the bag-of-words feature instead of the structured data. The construction of the feature vector is summarized in Figure 1.

The title, abstract, and main contents of the patent document are tokenized by a morphological analyzer.

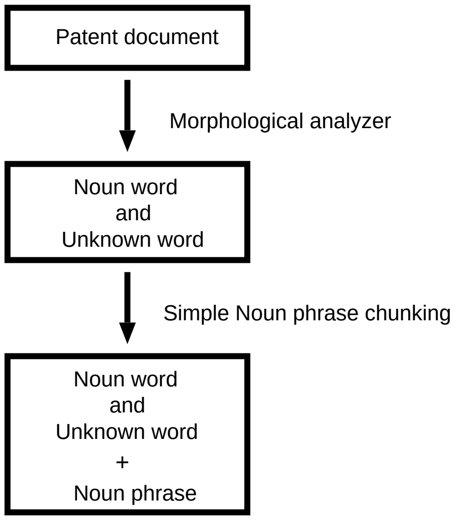

Figure 1. Process for feature selection

We used MeCab 0.92 [8] as the morphological analyzer. Next, noun words are extracted as features. These noun words are chunked by using very simple rules. These noun phrases are added to the feature vector. The value of each feature is set by a novel tf $\times$ idf value. Then, each feature vector corresponding to the patent document is normalized by a L2-norm. An example of a feature vector constructed by this procedure is shown in Figure 2.

```
PATENT-JA-UPA-1993-001033

   予防:0.0314
   反射-性-排尿-収縮-抑制-作用:0.0572
   低級-アルコキシ-基:0.0361
   薬理-的:0.2104
   エーテル:0.014
   メチル:0.0062
   これら:0.0006
   斉-炭素-原子:0.0199
   臭化-リチウム:0.0430
   カラメル:0.0445
   有機酸-付加-塩:0.0890
   機能:0.0136
```

Figure 2. Example of feature vector

where “-” represents the boundary of the tokenized term.

In a patent document, technical terms are frequently used. These terms are generally not included in the dictionary of the morphological analyzer. Thus, they may prevent the appropriate feature terms from being extracted. Meanwhile, the F-term definition file included in the patent map guidance system (PMGS.tgz) contains information regarding the classification of the tree structure and the description of F-terms. This information appears to be important for F-term classification. In particular, the keyword observed in the description of the F-term appears to be a good feature word for F-term classification. Therefore, we include the noun phrases extracted from the F-term definition file in the dictionary of the morphological analyzer. We think that this procedure will assist in the extraction of the appropriate feature terms.

The words shown in Figure 3 are a part of the extracted terms from theme 4C055 of the F-term definition file.

```
炭化水素
脂肪-族-炭化-水素
脂環-式-炭化-水素
芳香-族-炭化-水素
置換-炭化-水素
ハロゲン-置換-炭化-水素
酸素-原子-置換-炭化-水素
硫黄-原子-置換-炭化-水素
窒素-原子-置換-炭化-水素
```

Figure 3. Extracted words from F-term definition file

3 Classification scheme

3.1 Multi topic classifier

The multi-topic classifier was studied for several groups and formulations [10],[6],[9]. Based on these studies, we constructed the MTC possessing a loss function by means of the average precision of the ranked output F-terms.

Now, let $\mathbf{y} = (y_1, \ldots, y_l)$ be the feature vector associated with multiple F-terms, i.e., if $y_p = 1$, the p-th F-term is given in the document, if $y_p = 0$, p-th F-term is not given in the document. Let $\mathbf{x}$ be the bag-of-words representing the patent document. From this notation, we can determine the output F-term by using the following decision function:

$$\arg\max_{\mathbf{y}} \; score(\mathbf{y}, \mathbf{x}) = \sum_{p}^{l} y_p \langle \mathbf{w}_p, \mathbf{x} \rangle, \qquad (1)$$

where $\mathbf{w}_p$ is the weight vector associated with the p-th F-term and $\langle , \rangle$ is a novel inner product.

The set of weight vectors $\mathbf{w} = \{\mathbf{w}_1, \ldots, \mathbf{w}_l\}$ is determined by the following optimization problem:

$$\text{min.} \qquad \frac{1}{2}\sum_{p=1}^{l} \langle \mathbf{w}_p, \mathbf{w}_p \rangle \qquad (2)$$

$$\text{s.t.} \qquad margin_{\mathbf{w}}(\mathbf{y}^i, \mathbf{x}^i, \mathbf{y}') = \sum_{p=1}^{l} (y_p^i - y_p') \langle \mathbf{w}_p, \mathbf{x}^i \rangle$$

$$-\left\{1 - \text{avg. prec}(\mathbf{y}^i, \mathbf{y}^{'})\right\} \geq 0, \quad i = 1, \ldots, N, \tag{3}$$

where avg. prec$(\mathbf{y}^i, \mathbf{y}^{'})$ is the average precision between the i-th training multiple F-terms and the ranked estimated F-terms $\mathbf{y}^{'}$ [7].

We will explain the above expression. Because $\mathbf{w}_p$ is the weight vector and $\langle \mathbf{w}_p, \mathbf{x} \rangle$ is interpreted as the degree of the p-th F-term, we can naturally interpret equation (1) as the score of the given multiple F-terms. Equation (3) represents the F-terms that are correctly classified in the training data. Equation (2) represents the regularization term under the constraints of equation (3). This corresponds to the maximum margin principle [6].

To obtain a solution to the optimization problem of equations (2) and (3), we used the following update formula for $\{\mathbf{w}_p\}$.

$$\mathbf{w}_p^{(k+1)} = \mathbf{w}_p^{(k)} + c_k \left\{(y_p - y_p^{'})\mathbf{x}^i\right\}, \tag{4}$$

$$c_k = \max \left\{ z_i^{(k)}, \frac{-margin_{\mathbf{w}^{(k)}}(\mathbf{y}^i, \mathbf{x}^i, \mathbf{y}^{'})}{||\mathbf{y}^i - \mathbf{y}^{'}||^2 \, ||\mathbf{x}^i||^2} \right\}, \tag{5}$$

$$z_i^{(k+1)} = \begin{cases} z_i^{(k)} & \text{if} \quad z_i^{(k)} = c_k, \\ z_i^{(k)} - c_k & \text{otherwise.} \end{cases} \tag{6}$$

The training algorithm for MTC, which uses the updated formula for $\mathbf{w}$, is summarized as follows:

Training algorithm for the MTC

1. Initialize $k \leftarrow 0$, $\{\mathbf{w}_p^{(0)}\} = \mathbf{0}$, set MAX_ITER.
2. For each i, using the current $\mathbf{w}^{(k)}$ and decision function of equation (1), estimate the ranked output F-term $\mathbf{y}^{'}$.

   Then, update $\{\mathbf{w}_p^{(k)}\}$ using (4), (5) and (6)
3. If all the constraints of equation (3) are satisfied in step 2, or k reaches MAX_ITER, the iteration is terminated. otherwise, set $k \leftarrow k + 1$ and go to step 2

This update formula is almost the same as that used for a perceptron, and it corresponds to the approximate version of Hildreth's quadratic programming (QP) solution algorithm [2]. Our formulation does not rigorously satisfy the condition of Hildreth's method. However, we confirmed that the F-term is classified almost correctly in the training process.

3.2 One-vs-rest SVM classifier

The one-vs-rest SVM classifier is a combination of the binary SVM classifiers associated with all the F-terms [5]. The output score of each SVM represents the degree of F-term occurrence in a given patent document. We can determine the output F-terms using these scores.

The score is calculated by using a linear kernel function as follows:

$$score(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b. \tag{7}$$

Here, $\mathbf{w}, b$ are the weight vector and bias term, respectively; they are estimated by using a well known optimization problem [5].

The one-vs-rest SVM classifier requires almost the same training process for each F-term; this results in a large training time. In addition, the size of the negative examples is significantly greater than that of the positive ones; furthermore, almost all the negative examples other than the special examples, which are called as support vectors, are irrelevant of classification.

Based on these results, we used heuristics that randomly reduce 70–90% of the negative examples for each F-term. We confirmed that this improves the accuracy as well as reduces the training time of the preliminary experiments.

3.3 One-vs-rest Naive Bayes classifier

We used the one-vs-rest Naive Bayes classifier for themes 4D075 and 5B057. In the same manner as that of the one-vs-rest SVM classifier, we can determine the occurrence of the F-terms by using the following generative probability $P(y_p = 1|\mathbf{x})$:

$$P(y_p = 1|\mathbf{x}) =\propto \prod_j P(x_j|y_p = 1)P(y_p = 1), \tag{8}$$

where $P(\mathbf{x}_j|y_p = 1)$ represents the generative probability of term x_j based on the p-th F-term and $P(y_p = 1)$ is prior of the p-th F-term. These parameters can be quickly estimated by the maximum likelihood estimation in large training data.

4 Experiments

4.1 Experimental settings

In the classification subtask at NTCIR-6 Patent Retrieval Task, we performed two runs, namely, JSPAT1 and JSPAT2. The one-vs-rest SVM classifier is employed in JSPAT1, while MTC is used in JSPAT2. The MTC was not applied in some of the themes in JSPAT2. The MTC maintains the weight vector in the memory in both training and classification phases and requires large memory resources. Therefore, it is difficult to apply the MTC to the theme with large training data. Furthermore, the SVM classifier requires a relatively large training time. From the computational resource viewpoint, it is difficult to apply MTC to all the

runid	#SVM	#MTC	#NB
JSPAT1	106	0	2
JSPAT2	40	66	2

Table 1. Number of classifiers in each run

runid	MAP	R-Prec	Avg. F1
JSPAT1	43.81	40.32	30.38
JSPAT2	43.55	40.03	32.78

Table 2. Results for the entire experiments

themes. NB was used in the two theme tasks in both the runs. Table 1 shows a number of themes applied to the classification algorithm in each run.

We applied the two heuristics for SVM in some theme tasks. We summarize these as follows:

- The reduction of negative examples is used in theme 2C088 - 2H111
- Division of learning is applied for SVM in theme 4J002

We set the cost parameter C = 1.0 and use the linear kernel function for all the SVMs. The reduction ratio is set to 0.7 in themes 2C088–2H111 and to 0.9 in theme 4J002, that is, 70% and 90% of the negative examples are reduced in each of the cases. We set the number of divisions to 3 in theme 4J002. As mentioned above, we used the classification algorithm in each theme differently. Therefore, the results for the entire experiment do not accurately represent the performance of each classifier. Therefore, we describe the experimental results for each classifier and each theme task in the appendix.

4.2 Results for the entire experiments

Formal run results are shown in Table 2.

We reduced the negative examples of SVM in 15 themes. In 13 of these themes, the classification was not executed properly, as shown in the appendix. In some cases, the size of the positive examples is greater than that of the negative ones and the unbalancedness of the positive and negative examples increases due to this reduction. As a result, the classification did not work properly.

We show all the results for each classifier in Table 3. Here, the 13 SVM results that appear to be incorrect are excluded.

From Table 3, the SVM shows a relatively good MAP performance as compared to the best MAP run and the best F1 run. We describe the effects of reduction of negative examples, division of learning, and MTC in the following subsections.

algorithm	MAP	R-Prec	Avg. F1
SVM	47.38	43.55	33.22
MTC	43.69	39.47	33.64
NB	39.25	36.68	31.05
best MAP run	48.52	43.14	40.37
best F1 run	47.79	43.63	41.25

Table 3. Classification results for each algorithm

runid	MAP	R-Prec	Avg. F1
best MAP run	55.44	49.60	40.96
best F1 run	54.94	50.11	48.66
JSPAT1(SVM)	55.34	50.57	47.83

Table 4. Classification results for 2H079

4.3 The reduction of negative examples

We show the classification results for themes 2H079 and 2H111. In these themes, the reduction of negative examples is employed and 70% of the negative examples are reduced randomly.

In general, SVM shows good precision; however, the F1 results are not so good. However, both the results shown in Tables 4 and 5 represent a good performance as compared to the other runs. In particular, JSPAT1 shows the best MAP and F1 scores in 2H111 task. This shows the effectiveness of the reduction of negative examples under appropriate parameter settings. From the training time viewpoint, the reduction is effective, although we did not make rigorous comparisons .

4.4 Division of learning for SVM

Theme 4J002 has the biggest training data in this task; the size of the F-term set is 710 and that of the training document is 35147. We could not execute training directly for all the classifiers. Therefore, the division of learning and reduction of negative examples are applied for the one-vs-rest SVM.

As shown in Table 6, both the heuristics work well. We obtain the best F1 score in this theme. This indicates that ensemble learning when the division of learning is carried out.

4.5 Multi topic classifier

We developed a new MTC for this task. In developing, some formulation, for example, selection of

runid	MAP	R-Prec	Avg. F1
best MAP run	50.21	44.60	36.61
best F1 run	54.17	51.00	45.52
JSPAT1(SVM)	54.27	49.48	45.86

Table 5. Classification results for 2H111

runid	MAP	R-Prec	Avg. F1
best MAP run	40.48	39.31	24.24
best F1 run	44.55	44.60	30.74
JSPAT1(SVM)	42.91	41.96	36.57

Table 6. Classification results for 4J002

loss function, introduction of the similarity for each F-term, selection of the solution algorithm for QP problem is tried. However, the development of the classifier is not perfect. Therefore, the results of the MTC do not represent the effect of multi-topic classification in all the results. However, in some themes, the MTC shows relatively good performance, as shown in Table 5. Furthermore, the MTC outperforms the normal SVM with regard to the F1 score in many themes. (see Appendix)

5 Conclusion

We conducted the classification subtask at NTCIR-6 Patent Retrieval Task by three document classifiers, namely, the one-vs-rest SVM classifier, multi-topic classifier, and one-vs-rest Naive Bayes classifier.

For the one-vs-rest SVM classifier, two heuristics, namely, the reduction of negative examples and division of learning, are applied for the some of the themes. These heuristics improve the F1 score and reduced the training time in the three themes. However, the reduction of negative examples created a problem for the classification in the 13 themes. The cause of this problem appears to be the violation of our assumption regarding negative examples. We need to study and develop a safer and more general reduction technique.

Considering that the F-term categorization subtask is a multi-topic classification task, we developed a multi-topic classifier based on the maximum margin principle. This classifier showed a slightly better performance with regard to the F1 score than the SVM; however, the improvement is not as significant as expected. Many formulations for multi-topic classification are being studied. We should continue to study multi-topic classification for tasks having many class categories with structures that are the same as those of the F-term classification.

During this task, we attempted to introduce some similarity between each F-term based on the F-term

runid	MAP	R-Prec	Avg. F1
best MAP run	46.16	36.73	29.00
best F1 run	45.07	36.82	33.48
JSPAT1(SVM)	45.13	36.82	27.42
JSPAT2(MTC)	45.62	37.11	30.11

Table 7. Classification results for 4D040

classification system and their frequency of appearance of the F-terms [3]. However, positive results were not obtained in at least some formulations. We think that the introduction of this type of similarity can improve the classification performance; we are going to study the formulation of the similarity between the class categories such as F-terms.

References

[1] A. Fujii, M. Iwayama, and N. Kando. Overview of the Patent Retrieval Task at the NTCIR-6 Workshop. In *Proceedings of the Sixth NTCIR Workshop Meeting.*, 2007.

[2] C. Hildreth. A quadratic programming procedure. *Naval Research Logistics Quarterly*, 4:79–85, 1957.

[3] H. Hirai, K. Murota, and M. Rikitoku. SVM kernel by electric network. *Pacifi c Journal of Optimization*, 1(3):509–526, 2005.

[4] H.-Y. J. Jae-Ho Kim, Jin-Xia Huang and K.-S. ChoiT. Patent Document Retrieval and Classifi cation at KAIST. In *The proceedings of the Fifth NTCIR Workshop Meeting on Evaluation of Information access and Technologies*, 2005.

[5] T. Joachims. Text categorization with support vector machines: learning with many relevant features. In C. Nédellec and C. Rouveirol, editors, *Proceedings of ECML-98, 10th European Conference on Machine Learning*, pages 137–142, Chemnitz, DE, 1998. Springer Verlag, Heidelberg, DE.

[6] H. Kazawa, T. Izumitani, H. Taira, and E. Maeda. Maximal Margin Labeling for Multi-Topic Text Categorization. *Advances in Neural Information Processing Systems 17*, pages 649–656, 2005.

[7] R. McDonald, K. Crammer, and F. Pereira. Online Large-Margin Training of Dependency Parsers. *Ann Arbor*, 100, 2005.

[8] T. Kudo. MeCab: Yet Another Part-of-Speech and Morphological Analyzer. `http://mecab.sourceforge.net/`.

[9] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support vector machine learning for interdependent and structured output spaces. *ACM International Conference Proceeding Series*, 2004.

[10] N. Ueda and K. Saito. Parametric mixture models for multi-labeled text. *Advances in Neural Information Processing Systems*, 15:649–656, 2003.

6 Appendix: Classification results for all themes

Theme	MAP	R-Prec	F1
2C088	4.29	3.82	2.27
2D040	26.79	26.39	19.57
2D055	14.54	13.04	10.66
2E164	18.56	14.25	10.38
2F014	19.24	12.93	8.21
2F062	12.24	16.44	4.64
2F065	12.04	15.54	4.74
2F112	24.75	22.42	7.01
2G051	21.16	22.48	8.83
2G065	17.81	18.19	7.58
2H005	23.91	21.83	6.34
2H023	14.62	11.71	8.08
2H026	30.8	27.22	27.09
2H079	55.34	50.57	47.83
2H111	54.27	49.48	45.86

Table 8. Classification results of SVM for 2C088 - 2H111

Theme	MAP	R-Prec	Avg. F1
2D040	53.04	48.73	43.41
2D055	46.46	39.5	34.6
2E164	42.58	35.09	26.7
2F014	57.18	48.46	42.27
2F062	24.43	26.26	18.44
2F112	47.98	43.44	38.52
2H026	56.24	49.67	46.11
2H079	54.62	50.49	44.06

Table 9. Classification results of MTC for 2D040 - 2H079

Theme	MAP	R-Prec	Avg. F1
3B154	44.4	44.52	27.22
3C034	45.59	40.84	30.31
3C045	40.53	34.3	25.31
3D041	48.1	45.85	38.98
3D043	40.97	39.14	23.46
3D054	46.72	43.41	32.25
3E040	51.29	46.52	34.43
3E083	46.8	42.87	29.02
3F022	40.48	37.39	27.6
3F049	41.92	36.41	33.36
3F054	47.43	45.43	36.36
3F064	54.25	48.26	35.67
3F079	42.92	40.57	29.11
3F301	58.1	50.33	40.08
3F307	49.74	43.07	32.72
3F333	46.97	44.82	36.5
3G019	51.44	48.68	30.05
3G023	51.03	47.35	38.34
3G091	49.68	47.93	36.85
3H045	49.57	45.98	35.94
3H076	43.41	38.48	25.6
3K068	38.38	35.09	23.83
3L050	55.66	44.2	39.23
3L051	47.67	38.51	30.59
3L103	47.58	43.75	29.17

Table 10. Classification results of SVM for 3B154 - 3L103

Theme	MAP	R-Prec	Avg. F1
3B154	42.21	41.25	33.42
3D041	45.07	43.68	40.11
3D043	39.76	38.12	27.39
3D054	42.78	40.01	32.6
3E040	46.37	41.19	34.68
3E083	46.09	40.69	32.65
3F022	36.32	33.19	28.6
3F049	40.35	35.07	34.47
3F064	51.07	44.64	38.69
3F079	40.13	36.73	32.45
3F307	46.82	41.61	33.26
3G019	48.67	47.38	33.61
3G023	47.9	42.91	40.51
3G091	42.14	42.85	37.18
3H045	46.89	43.97	36.7
3H076	43.05	38.77	32.7
3K068	36.55	32.93	28.03
3L050	55.57	46.06	40.29
3L051	1.57	1.57	2.03
3L103	43.95	41.16	35.74

Table 11. Classification results of MTC for 3B154 - 3L103

Theme	MAP	R-Prec	Avg. F1
4B017	51.62	45.98	27.3
4C055	49.36	46.14	39.81
4C063	57.35	51.2	44.38
4C084	46.53	44.33	39.48
4C090	45.62	43.86	26.9
4C093	39.77	39.22	21.92
4D012	47.99	42.37	29.1
4D040	45.13	36.82	27.42
4D050	54.79	50.43	40.87
4D059	35.32	35.42	17.33
4D065	45.71	41.83	29.54
4D075	34.49	34.47	27.68
4E081	31.85	30.05	16.33
4F070	38.16	37.74	24.03
4F071	53.67	51.67	46.38
4F073	45.27	42.74	30.2
4F210	55.67	54.25	49.58
4G072	49.0	47.79	32.57
4H045	59.26	54.2	47.87
4H057	61.34	57.24	47.61
4H061	55.49	52.37	36.52
4J002	42.91	41.96	36.57
4J034	26.19	32.65	27.37
4J039	52.37	50.94	31.83
4J043	45.01	45.96	44.69
4K001	49.98	46.71	39.35
4K013	54.18	49.22	41.27
4K026	59.62	55.13	41.62
4K031	46.63	42.49	30.07
4K044	60.41	55.34	47.5
4L045	63.48	57.15	49.88
4L056	39.15	37.59	24.7
4M112	54.44	49.28	41.18
4M118	52.87	49.92	39.92

Table 12. Classification results of SVM for 4B017 - 4M118

Theme	MAP	R-Prec	Avg. F1
4B017	49.11	42.94	35.19
4C063	52.47	46.57	44.43
4C090	44.72	42.19	33.67
4D012	46.27	40.08	32.38
4D040	45.62	37.11	30.81
4D050	52.73	48.56	41.94

Table 13. Classification results of MTC for 4B017 - 4D050

Theme	MAP	R-Prec	Avg. F1
5B013	46.17	35.66	28.76
5B029	56.56	48.54	45.95
5B034	43.19	35.24	28.45
5B057	44.16	38.96	34.53
5B062	46.87	40.25	27.92
5B064	40.85	38.75	34.18
5B076	36.35	30.0	25.45
5C023	51.83	46.24	32.47
5C055	53.51	48.85	38.96
5C060	37.58	33.49	23.65
5C087	39.39	39.36	23.75
5D015	45.61	38.05	33.5
5D042	49.68	42.86	32.38
5D046	33.11	30.16	17.09
5D117	42.08	36.99	29.81
5E077	46.59	43.18	35.32
5E082	51.82	50.43	42.79
5E319	48.73	42.96	37.65
5E346	38.63	39.34	34.97
5F051	43.62	39.5	32.59
5F056	34.82	32.11	18.04
5F101	42.6	40.57	25.17
5F102	61.59	57.05	49.51
5G321	51.56	45.91	32.77
5H007	55.75	50.89	35.93
5H024	49.84	46.36	31.73
5H030	49.99	45.37	28.5
5J065	55.32	52.31	47.0
5K024	45.24	41.44	29.6
5K026	32.74	29.59	19.61
5K039	41.59	37.09	23.83
5K051	37.23	34.55	22.79
5K061	48.04	43.97	30.91
5K072	45.59	41.93	27.11

Table 14. Classification results of SVM for 5B013 - 5K072

Theme	MAP	R-Prec	Avg. F1
5B013	44.29	33.64	28.17
5B029	54.74	47.13	45.86
5B034	42.6	32.87	29.99
5B062	44.15	37.74	30.01
5B064	37.44	35.77	31.81
5C023	49.62	43.57	38.17
5C055	51.76	48.38	41.31
5C060	35.05	30.32	25.65
5C087	37.09	37.26	28.28
5D015	42.61	34.96	32.75
5D042	48.68	41.32	33.84
5D046	29.05	26.4	19.48
5D117	39.63	35.34	29.69
5E077	43.18	40.1	36.83
5E082	45.97	45.82	41.68
5E319	44.5	38.75	37.26
5E346	37.03	36.25	34.43
5F051	38.87	34.57	32.17
5F056	34.36	31.15	21.94
5F101	39.11	37.28	28.39
5F102	58.74	54.88	50.42
5G321	49.57	43.75	36.44
5H007	50.3	46.35	38.74
5H024	48.49	44.56	36.6
5H030	35.04	36.39	24.58
5J065	53.74	50.31	47.22
5K024	42.3	38.6	32.79
5K026	32.73	29.74	26.43
5K039	38.35	34.04	25.25
5K051	35.45	31.67	27.23
5K061	46.25	41.09	34.44
5K072	40.77	38.13	31.0

Table 15. Classification results of MTC for 5B013 - 5K072

Theme	MAP	R-Prec	Avg. F1
4D075	34.49	34.47	27.68
5B057	44.16	38.96	34.53

Table 16. Classification results of NB for 4D075, 5B057