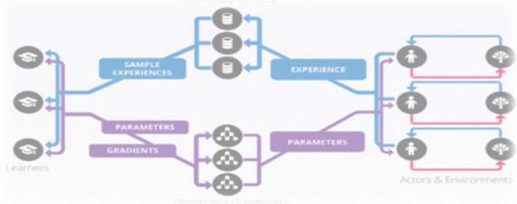


제 132회 석사 학위 졸업 논문 발표

Effective Multi-label Text Categorization using Pairwise Topic Dependency

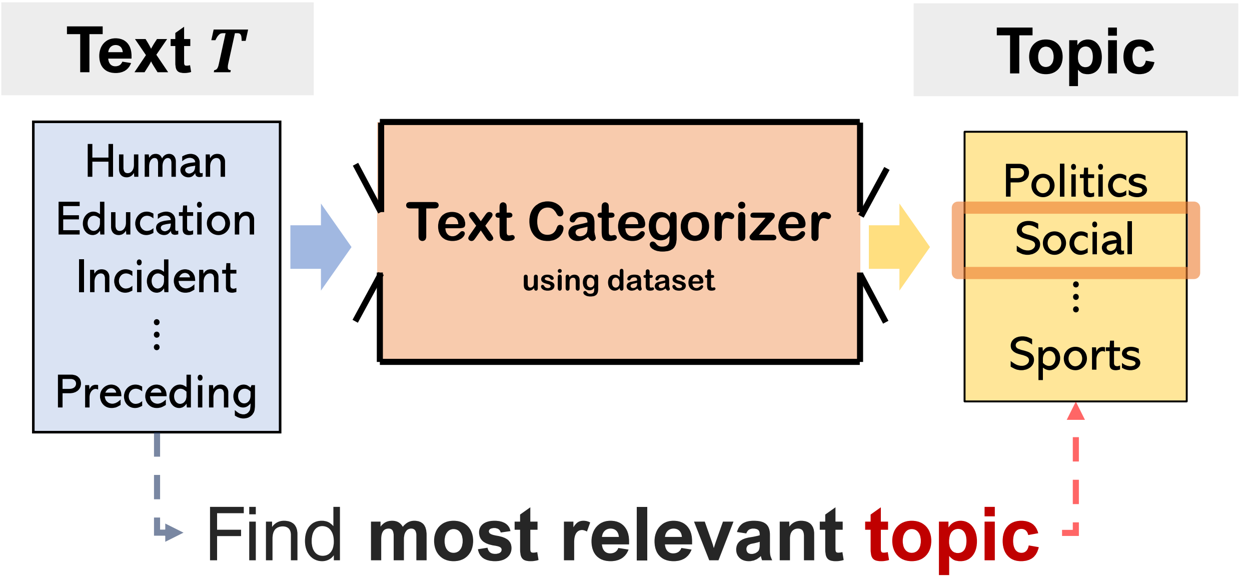
쌍방 주제 종속성을 이용한 효과적인 다중 레이블 텍스트 분류에 관한 연구

Hae-Cheon Kim
code.bug.station@gmail.com

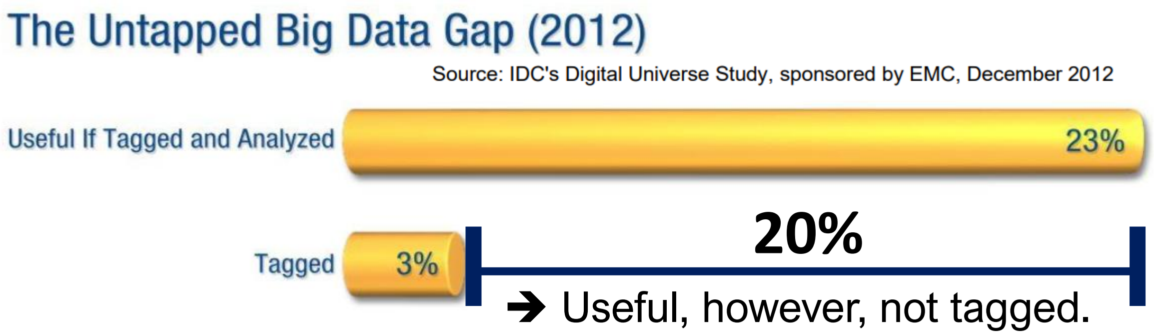
2019. 11. 22.

Text Categorization

- **Text Categorization**
  - Given **text T**,
    - Estimate the most relevant **topic of T**.



- In the digital internet world [1],
  - There are **80%** of data is formed by text.
  - **Text categorization** will makes 20% text data **available that can't use now**.
- Therefore, **Text categorization** is **important issue** in modern.



Naïve Bayes Classifier

- **Naïve Bayes classifier** has **good performance** in **text categorization** [2, 8, 9, 10].
- **Naïve Bayes classifier** calculates **posterior probability** based on two supposes.

Text x

Human
Education
Incident
⋮
Preceding

The Goal of Naïve Bayes Classifier

Find $\underset{y}{\arg\max}\ P(y|x)$

Bayes' theorem

$$P(y|x) = \frac{P(y) \cdot P(x|y)}{P(x)}$$

+

Naïve Independence Assumption

$$P(x|y) = \prod_{\forall i} P(x_i|y)$$

$$P(y|x) \propto P(y) \cdot \prod_{\forall i} P(x_i|y)$$

(posterior probability)

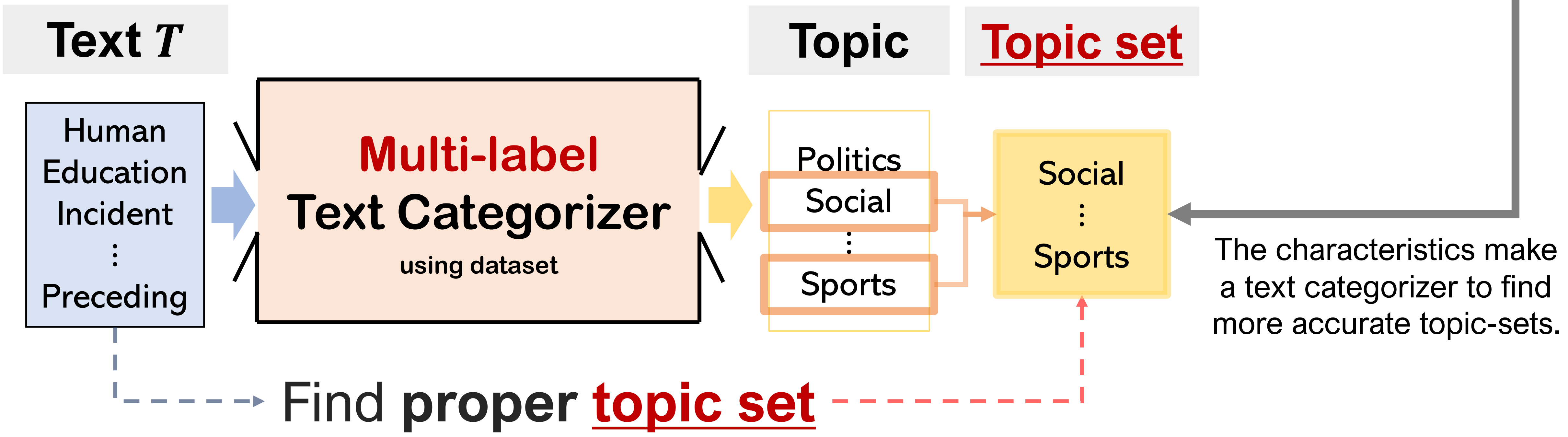
Topic y

Politics
Social
⋮
Sports

Multi-label Text Categorization

- **Multi-label Text Categorization**
  - Given **text T**,
    - Estimate the most proper **topic-set of T**.
- The generalized problem of text categorization.

- **Characteristic of Multi-label**
  - A Instance has many categories.
  - The **realistic topic-sets** are limited.
  - Some particular **topic-sets often appear frequently**.



Multi-label Naïve Bayes Text Categorization

- **Multi-label Text Categorization using Naïve Bayes Classifier**
  - Naïve Bayes classifier is widely used and it has good performance in text categorization.

The Goal of Multi-label Naïve Bayes Text Categorization

Find $\arg\max_{Y} P(Y|x)$

x: Input text
Y: Output topic-set

+

- **Characteristic of Multi-label**
  - A Instance has many categories.
  - The **realistic topic-sets** are limited.
  - Some particular **topic-sets often appear frequently**.

How to estimate the $P(Y|x)$ properly applying characteristic of multi-label?

Many Approaches of Multi-label Naïve Bayes Text Categorization

$$\arg\max_{Y} P(Y|x)$$

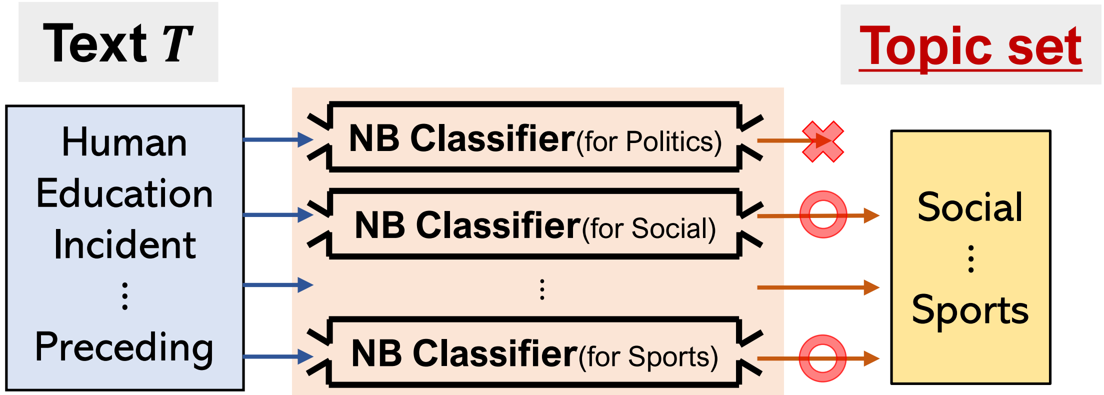
1. Calculate all the probability of all topic-sets

- The number of all possible topic-sets
  - $O(2^q)$ where q is the number of topics ➔ **Heavy time complexity**

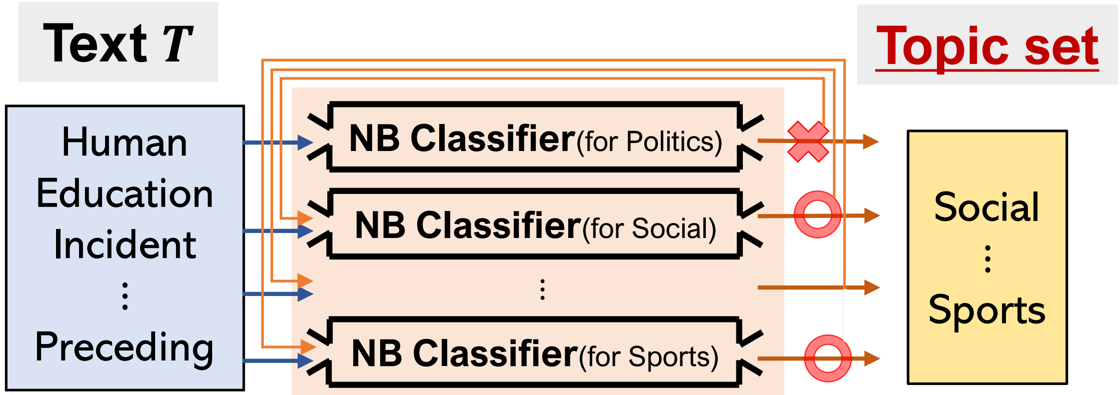
2. Calculate the result for each topic (MLNB [3]) ➔ **Ignore characteristic of multi-label**

3. Ensemble of classifier chains (ECC-NB [4])

- Set label order, then reflect the result of the previous label to the next label prediction.
  - Accuracy is affected by label order ➔ **Unstable probability**



- The structure of MLNB -



- The structure of ECC-NB -

Proposed Method

- Based on Bayes' theorem, I derive novel posterior probability equation.
- Derivation of proposed
  - Let $\mathbf{x}$ be a given data, $\mathbf{y}$ be a topic vector.
  - Assume the $\mathbf{x_1}, \ldots, \mathbf{x_d}, \mathbf{y_1}, \ldots, \mathbf{y_q}$ are conditional independently from $\mathbf{y_k}$.

$$\underset{\mathbf{y}\in[0,1]^q}{\arg\max}\, P(\mathbf{y}|\mathbf{x}) \propto P(\mathbf{x},\mathbf{y}) = P(\mathbf{y}_k)\cdot P(\mathbf{x}_1,\ldots,\mathbf{x}_d,\mathbf{y}_1,\ldots,\mathbf{y}_q|\mathbf{y}_k)$$

  - With Naïve Independence Assumption

$$\approx P(\mathbf{y}_k)\cdot P(\mathbf{x}_1,\ldots,\mathbf{x}_d|\mathbf{y}_k)\cdot P(\mathbf{y}_1,\ldots,\mathbf{y}_d|\mathbf{y}_k)$$

$$= P(\mathbf{y}_k)\prod_{i=1}^{d} P(\mathbf{x}_i|\mathbf{y}_k)\boxed{\prod_{j=1}^{q} P(\mathbf{y}_j|\mathbf{y}_k)}$$

  - The formula is based on $\mathbf{y_k}$. then calculate all probabilities for all topics and averages them.
  - Using a characteristic of multi-label, the realistic topic-sets are limited.

Pairwise Topic Dependency

$$\underset{\mathbf{y}\in\mathcal{D}_{\mathbf{y}}}{\arg\max}\, P(\mathbf{y}|\mathbf{x}) \propto \prod_{k=1}^{q}\left(P(\mathbf{y}_k)\prod_{i=1}^{d} P(\mathbf{x}_i|\mathbf{y}_k)\boxed{\prod_{j=1}^{q} P(\mathbf{y}_j|\mathbf{y}_k)}\right)$$

Experiment Setting

- Datasets: **11 Multi-Label Text Datasets**: Yahoo-dataset [7]

Name	Patterns	Features	Labels	Cardinality	Density	#Distinct label-set
Arts	7,484	1,157	26	1.654	0.064	599
Business	11,214	1,096	30	1.599	0.053	233
Computers	12,444	1,705	33	1.507	0.046	428
Education	12,030	1,377	33	1.463	0.044	511
Entertainment	12,730	1,600	21	1.414	0.067	337
Health	9,205	1,530	32	1.644	0.051	335
Recreation	12,828	1,516	22	1.429	0.065	530
Reference	8,027	1,984	33	1.174	0.036	275
Science	6,428	1,859	40	1.450	0.036	457
Social	12,111	2,618	39	1.279	0.033	361
Society	14,512	1,590	27	1.670	0.062	1,054

Experiment Setting

- Comparison Classifiers:

Name	Explanation	Hyper-parameter	Ref.
MLNB	Multi-label Naïve Bayes	-	[3] *Information Sciences*. (**2009**)
ECC-NB	Ensemble of Classifier Chains using Naïve Bayes	Ensemble 10 times	[4] *Machine learning*. (**2011**)
GLOCAL	Global and Local Label Correlation with Matrix Factorization	Threshold $= 0.5$, $k = [5, 10, \ldots, 25]$, $\lambda = 1$	[5] *IEEE TKDE*. (**2017**)
ML-kELM	Multi-label Kernel Extreme Learning Machine	$\sigma = 2^{-2}$, $C = [2^0, 2^1, 2^2, 2^3]$	[6] *Neurocomputing*. (**2017**)

  - Probability Model: Multinomial. Add-one(Laplace) Smoothing.

- Evaluation Measure:
  - Macro F_1 Score: Are the labels for each data correctly estimated?
  - Micro F_1 Score: Are each label properly estimated with the data?
  - Multi-label Accuracy: Are labels correctly estimated for each piece of data?

- Cross-validation:
  - **Holdout** (Training: 80%, Testing: 20%), Repeated 10 times

Experiment Results (Macro F_1)

Table 4.2: Comparison results in terms of Macro F_1

Dataset	Proposed	MLNB	ECC-NB	ML-kELM	GLOCAL
Arts	**0.3501 ± 0.0283**✓	0.2768 ± 0.0081	0.2781 ± 0.0083	0.1463 ± 0.0100	0.0569 ± 0.0219
Business	**0.3497 ± 0.0339**✓	0.2502 ± 0.0096	0.2461 ± 0.0071	0.0858 ± 0.0009	0.0737 ± 0.0312
Computer	**0.2992 ± 0.0224**✓	0.2535 ± 0.0104	0.2469 ± 0.0096	0.0843 ± 0.0014	0.0847 ± 0.0347
Education	**0.3052 ± 0.0306**✓	0.2296 ± 0.0053	0.2241 ± 0.0053	0.1387 ± 0.0117	0.0594 ± 0.0190
Entertainment	**0.3328 ± 0.0135**✓	0.3085 ± 0.0064	0.3036 ± 0.0058	0.1853 ± 0.0072	0.0968 ± 0.0204
Health	**0.4094 ± 0.0319**✓	0.2603 ± 0.0069	0.2591 ± 0.0069	0.1805 ± 0.0122	0.1434 ± 0.0208
Recreation	**0.3945 ± 0.0145**✓	0.3591 ± 0.0100	0.3634 ± 0.0107	0.2249 ± 0.0073	0.0790 ± 0.0208
Reference	**0.3427 ± 0.0520**✓	0.2040 ± 0.0104	0.2029 ± 0.0087	0.0875 ± 0.0037	0.0720 ± 0.0262
Science	**0.3085 ± 0.0203**✓	0.2446 ± 0.0083	0.2428 ± 0.0075	0.0854 ± 0.0045	0.0816 ± 0.0540
Social	**0.2688 ± 0.0256**✓	0.1633 ± 0.0091	0.1552 ± 0.0098	0.0941 ± 0.0031	0.0404 ± 0.0039
Society	**0.2725 ± 0.0149**✓	0.2029 ± 0.0033	0.1996 ± 0.0026	0.1188 ± 0.0054	0.0312 ± 0.0099
Avg. Rank.	**1.000**	2.188	2.818	4.0909	4.9090

Experiment Results (Micro F_1)

Table 4.4: Comparison results in terms of Micro F_1

Dataset	Proposed	MLNB	ECC-NB	ML-kELM	GLOCAL
Arts	**0.4313 ± 0.0085**✓	0.3920 ± 0.0079	0.3918 ± 0.0079	0.2583 ± 0.0108	0.1555 ± 0.0549
Business	0.6529 ± 0.0052	0.6102 ± 0.0069	0.6007 ± 0.0064	0.1525 ± 0.0018	**0.7090 ± 0.0095**✓
Computer	0.4948 ± 0.0047	0.4473 ± 0.0074	0.4370 ± 0.0069	0.1091 ± 0.0017	**0.5004 ± 0.0027**✓
Education	**0.4409 ± 0.0068**✓	0.4052 ± 0.0065	0.3998 ± 0.0070	0.3157 ± 0.0074	0.2542 ± 0.0802
Entertainment	**0.5009 ± 0.0075**✓	0.4597 ± 0.0046	0.4528 ± 0.0055	0.3233 ± 0.0087	0.2610 ± 0.0442
Health	**0.5843 ± 0.0033**✓	0.5639 ± 0.0054	0.5604 ± 0.0065	0.4556 ± 0.0130	0.5207 ± 0.0661
Recreation	**0.4579 ± 0.0060**✓	0.4414 ± 0.0087	0.4452 ± 0.0087	0.2974 ± 0.0070	0.1434 ± 0.0377
Reference	**0.5086 ± 0.0084**✓	0.4478 ± 0.0083	0.4438 ± 0.0102	0.2666 ± 0.0125	0.4231 ± 0.0864
Science	**0.4151 ± 0.0082**✓	0.3585 ± 0.0073	0.3526 ± 0.0078	0.1588 ± 0.0080	0.2847 ± 0.1113
Social	**0.5121 ± 0.0110**✓	0.3481 ± 0.0236	0.3102 ± 0.0309	0.3143 ± 0.0093	0.4559 ± 0.0429
Society	**0.3650 ± 0.0050**✓	0.2941 ± 0.0074	0.2748 ± 0.0090	0.2704 ± 0.0061	0.2298 ± 0.0410
Avg. Rank.	**1.182**	2.363	3.272	4.454	3.727

Experiment Results (Multi-label Accuracy)

Table 4.6: Comparison results in terms of Multi-label accuracy

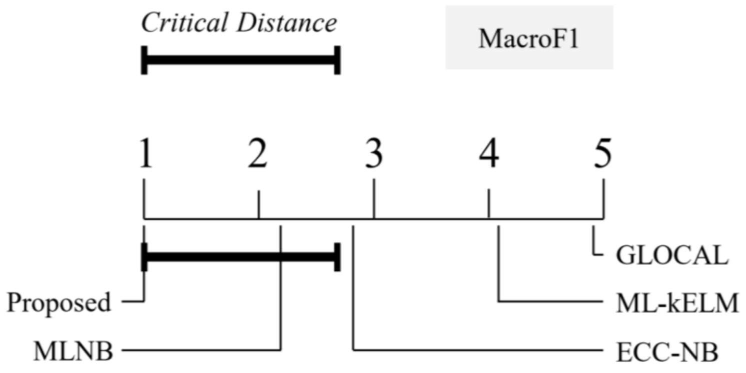
Dataset	Proposed	MLNB	ECC-NB	ML-kELM	GLOCAL
Arts	**0.3489 ± 0.0085**✓	0.3305 ± 0.0063	0.3344 ± 0.0068	0.2218 ± 0.0096	0.1055 ± 0.0403
Business	0.5858 ± 0.0048	0.5898 ± 0.0047	0.5885 ± 0.0042	0.2150 ± 0.0141	**0.7030 ± 0.0794**✓
Computer	0.4074 ± 0.0069	0.3966 ± 0.0079	0.3925 ± 0.0080	0.1028 ± 0.0056	**0.4629 ± 0.0093**✓
Education	**0.3431 ± 0.0067**✓	0.3367 ± 0.0092	0.3392 ± 0.0095	0.2401 ± 0.0061	0.1691 ± 0.0065
Entertainment	**0.4537 ± 0.0070**✓	0.4416 ± 0.0054	0.4420 ± 0.0053	0.3034 ± 0.0073	0.1822 ± 0.0289
Health	0.4847 ± 0.0043	0.4987 ± 0.0052	**0.5019 ± 0.0048**✓	0.4384 ± 0.0111	0.4579 ± 0.0787
Recreation	0.3943 ± 0.0065	0.3941 ± 0.0087	**0.3990 ± 0.0083**✓	0.2608 ± 0.0055	0.0910 ± 0.0260
Reference	**0.4647 ± 0.0094**✓	0.4529 ± 0.0086	0.4573 ± 0.0090	0.3876 ± 0.0149	0.3250 ± 0.1026
Science	**0.3422 ± 0.0081**✓	0.3276 ± 0.0070	0.3289 ± 0.0084	0.1823 ± 0.0085	0.2088 ± 0.1012
Social	**0.5126 ± 0.0118**✓	0.4489 ± 0.0217	0.4445 ± 0.0242	0.4543 ± 0.0122	0.3643 ± 0.0473
Society	**0.2985 ± 0.0064**✓	0.2659 ± 0.0082	0.2583 ± 0.0087	0.2646 ± 0.0066	0.1839 ± 0.0372
Avg. Rank.	**1.636**	2.727	2.454	4.090	4.090

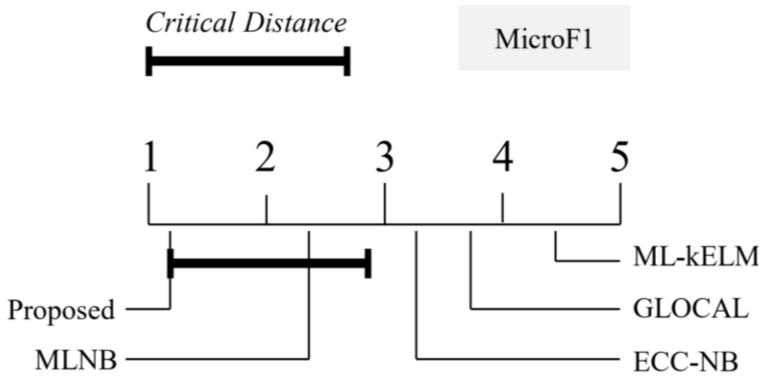
Experiment Results

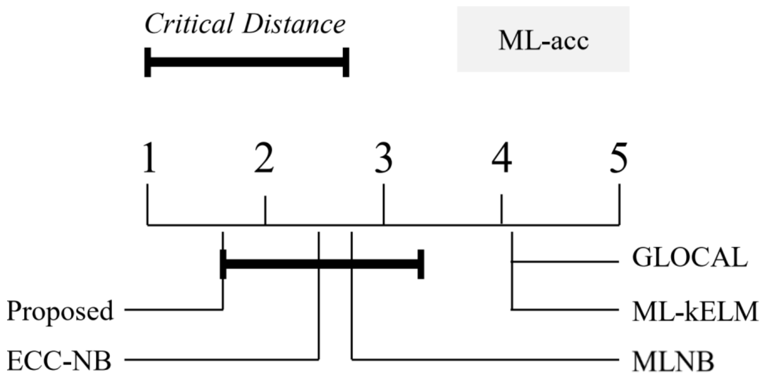
- Friedman Test $(\alpha = 0.95)$

Rank Avg.	Proposed	MLNB	ECC_NB	GLOCAL	ML-kELM	χ^2	$p = 0.05$
MacroF1	1.000	2.188	2.818	4.909	4.091	206.071	
MicroF1	1.182	2.363	3.272	3.727	4.454	18.009	2.4004
ML-ACC	1.636	2.727	2.454	4.090	4.090	8.558	

- Bonferroni-Dunn test $(\alpha = 0.95)$
  - Critical Distance: 1.6841







Experiment Results between Proposed and MLNB

- Wilcoxon's rank sum test ($p = 0.95$)

Table 4.9: The p-values of wilcoxon's rank sum test.

Dataset	Macro F_1	Micro F_1	Multi-label Accuracy
Arts	0.00018 ✓	0.00018 ✓	0.00032 ✓
Business	0.00018 ✓	0.00018 ✓	0.07570
Computers	0.00024 ✓	0.00018 ✓	0.01130 ✓
Education	0.00018 ✓	0.00018 ✓	0.27300
Entertainment	0.00058 ✓	0.00018 ✓	0.00170 ✓
Health	0.00018 ✓	0.00018 ✓	0.00018 ✓
Recreation	0.00032 ✓	0.00018 ✓	0.96980
Reference	0.00018 ✓	0.00018 ✓	0.02570 ✓
Science	0.00018 ✓	0.00018 ✓	0.00220 ✓
Social	0.00018 ✓	0.00018 ✓	0.00024 ✓
Society	0.00018 ✓	0.00018 ✓	0.00018 ✓

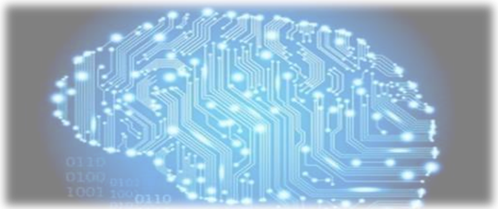
- ML-Acc of Proposed & MLNB

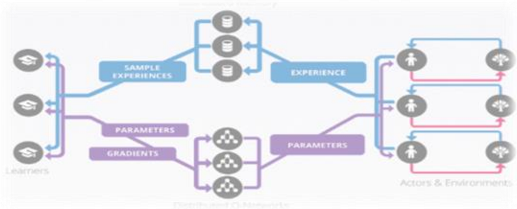
Dataset	Proposed	MLNB
Arts	**0.3489±0.0085✓**	0.3305±0.0063
Business	0.5858±0.0048	0.5898±0.0047
Computers	**0.4074±0.0069✓**	0.3966±0.0079
Education	0.3431±0.0067	0.3367±0.0092
Entertainment	**0.4537±0.0070✓**	0.4416±0.0054
Health	0.4847±0.0043	**0.4987±0.0052✓**
Recreation	0.3943±0.0065	0.3941±0.0087
Reference	**0.4647±0.0094✓**	0.4529±0.0086
Science	**0.3422±0.0081✓**	0.3276±0.0070
Social	**0.5126±0.0118✓**	0.4489±0.0217
Society	**0.2985±0.0064✓**	0.2659±0.0082

Conclusions & Contribution

- I propose "Effective Multi-label Naïve Bayes Text Categorization **with Pairwise Topic Dependency**".
- The approach reflects topic dependency with a novel posterior probability equation.
- The approach allowed us to estimate realistic topic-set more efficiently when attempting categorization.
- The experiments show that topic dependency can improve categorization accuracy.

- Contribution
  - In International Publications (SCI-E)
    - MLNB-LD: Multi-label Naive Bayes Classifier considering Label Dependency
    - ***Pattern Recognition Letter***, In submission, **1st author**
  - In Domestic Publications (KCI)
    - A Novel Posterior Probability Estimation Method for Multi-label Naive Bayes Classification
    - ***한국컴퓨터정보학회 논문집***, 1 June 2018, **1st author**





Thank You

Question & Answer

Reference

[1] Gantz, John, and David Reinsel. “The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east.” *IDC iView: IDC Analyze the future 2012*. (**2012**): 1-16.

[2] McCallum, Andrew, and Kamal Nigam. “A comparison of event models for naive Bayes text classification.” *AAAI-98 workshop on learning for text categorization*. Vol. 752. No. 1. (**1998**): 41-48.

[3] Zhang, Min-Ling, José M. Peña, and Victor Robles. “Feature selection for multi-label naive Bayes classification.” *Information Sciences*. Vol. 179. No. 19. (**2009**): 3218-3229.

[4] Read, Jesse, et al. “Classifier chains for multi-label classification.” *Machine learning*. Vol. 85. No. 3. (**2011**): 333-359.

[5] Zhu, Yue, James T. Kwok, and Zhi-Hua Zhou. “Multi-label learning with global and local label correlation.” *IEEE Transactions on Knowledge and Data Engineering*. Vol. 30. No. 6. (**2017**): 1081-1094.

[6] Luo, Fangfang, et al. “A multi-label classification algorithm based on kernel extreme learning machine.” *Neurocomputing*. Vol. 260. (**2017**): 313-320.

[7] Ueda, N. and Saito, K. “Parametric mixture models for multi-labeled text.” *Advances in Neural Information Processing Systems*. Vol. 15. (**2003**): 737–744.

Reference

[8] Gao, Hongyi, Xi Zeng, and Chunhua Yao. "Application of improved distributed naive Bayesian algorithms in text classification." *Journal of Supercomputing*. (**2019**): to be appeared. 1-17.

[9] Deng, Xuelian, et al. "Feature selection for text classification: A review." *Multimedia Tools and Applications*. Vol. 78. No. 3. (**2019**): 3797-3816.

[10] Xu, Shuo. "Bayesian Naïve Bayes classifiers to text classification." *Journal of Information Science*. Vol. 44. No. 1. (**2018**): 48-59.

Appendix. Contributions in Master Course

- In SCI-E Journal,
  - MLNB-LD: Multi-label Naive Bayes Classifier considering Label Dependency
    - ***Pattern Recognition Letter***, 19 June 2019, **1st author**
  - Competitive Particle Swarm Optimization for Multi-Category Text Feature Selection
    - ***Entropy***, 19 June 2019, **Co-author**

- In KCI Journal,
  - A Novel Posterior Probability Estimation Method for Multi-label Naive Bayes Classification
    - ***한국컴퓨터정보학회 논문집***, 1 June 2018, **1st author**

- In Domestic Conferences,
  - Three papers were submitted on the theme of Naive Bayes, Feature Selection, and Path Planning.