The performance of BERT as data representation of text clustering - Journal of Big Data 2022

저작권디지털포렌식전공 석사과정

2021021635 박재하

Related Work for 문서파일 분류

1. 문서파일 확장자 텍스트 추출
   - 참조 논문 : 유병영, 이상진, "디지털 포렌식 조사를 위한 문서 필터 도구 개발", 고려대학교 석사논문, 2011

2. **BERT를 활용한 주제별 문서 분류**
   - **참조 논문 : Alvin Subakti et al, "The performance of BERT as data representation of text clustering", Journal of Big Data, 2022**

논문 개요

- 본 논문은 **Text Clustering**에서 가장 널리 사용되는 **TF-IDF** 알고리즘의 한계를 설명하고, 이를 개선하기 위해 다양한 NLP Task에서 SOTA(State-of-the-art)를 달성한 **BERT**를 활용한 Clustering의 성능을 테스트함.
    - ML 분야에서 Classification은 Training Data를 활용한 지도 학습(Supervised Learning)이며, 훈련 없이 결과를 도출하는 비지도 학습(Unsupervised Learning)에 해당하는 문서 분류는 Clustering에 해당함

- TF-IDF는 검색엔진 등 여러 NLP Task에서 여전히 유용하지만, 단어의 (문장 내에서의) 맥락을 고려하지 못하므로, 이것이 가능한 BERT가 더 뛰어날 것으로 판단, 실제 실험 결과 많은 케이스에서 더 뛰어난 성능을 보임.

TF-IDF

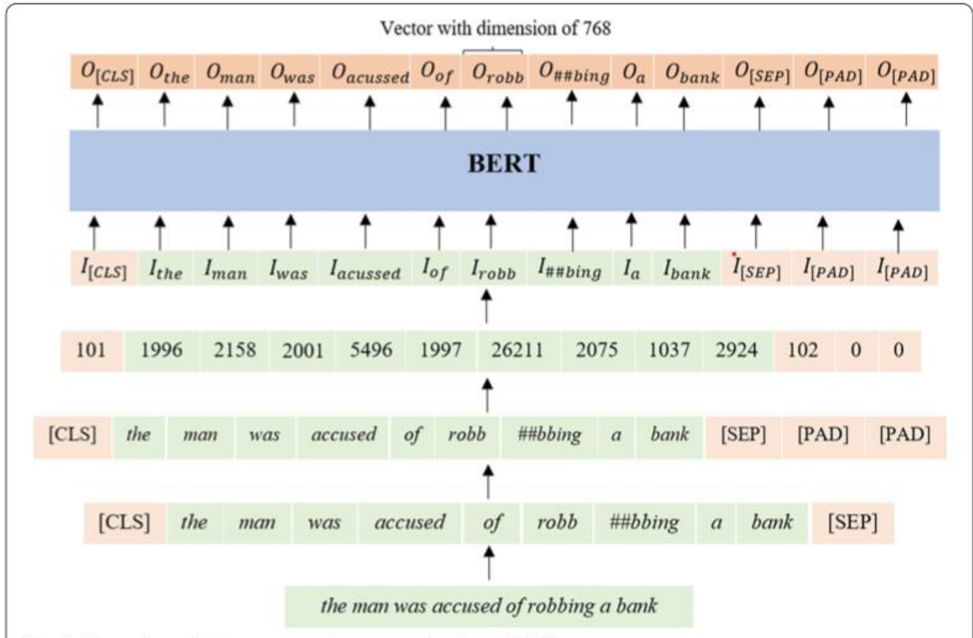
- Term Frequency-Inverse Document Frequency
  - tf : 문서 ***d***에서 단어 ***t***의 빈도
  - df : 전체 문서에서 단어 ***t***의 빈도
  - N : 전체 문서의 수
  - w : 단어 ***t***의 TF-IDF 값
- 어디에나 나오는 단어면 d에서 많이나와도 낮은값, 희소한 단어인데 d에서 많이 나와야 높은값

$$w_{t,d} = tf_{t,d} \times \log\left(\frac{N}{df_t}\right),$$

BERT

- Bidirectional Encoder Representation from Transformers
- Contextualized Word Embedding의 일종
    - 문장 단위로 단어 임베딩을 수행 (같은 단어라도 문맥에 따라 다르게 임베딩)

- Feature extraction
    - 768 dimension의 feature를 Max/Mean pooling



Vector with dimension of 768
$O_{[CLS]}$ O_{the} O_{man} O_{was} $O_{acussed}$ O_{of} O_{robb} $O_{\#\#bing}$ O_{a} O_{bank} $O_{[SEP]}$ $O_{[PAD]}$ $O_{[PAD]}$
BERT
$I_{[CLS]}$ I_{the} I_{man} I_{was} $I_{acussed}$ I_{of} I_{robb} $I_{\#\#bing}$ I_{a} I_{bank} $I_{[SEP]}$ $I_{[PAD]}$ $I_{[PAD]}$
101 1996 2158 2001 5496 1997 26211 2075 1037 2924 102 0 0
[CLS] the man was accused of robb ##bbing a bank [SEP] [PAD] [PAD]
[CLS] the man was accused of robb ##bbing a bank [SEP]
the man was accused of robbing a bank

Fig. 2 Illustration of text representation extraction from BERT

KM, EFCM, DEC, IDEC

$$r_{nk} = \begin{cases} 1, & k = \arg\min_k ||\boldsymbol{x}_n - \mu_k||^2 \\ 0, & \text{others} \end{cases}$$

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} ||\boldsymbol{x}_n - \mu_k||^2. \quad \mu_k = \frac{\sum_{n=1}^{N} r_{nk} \boldsymbol{x}_n}{\sum_{n=1}^{N} r_{nk}}.$$

- K-Means clustering
- Eigenspace-based Fuzzy C-Means

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk}^{m} ||\boldsymbol{x}_n - \mu_k||^2. \quad r_{nk} = \frac{1}{\sum_{j=1}^{K} \left(\frac{||\boldsymbol{x}_n - \mu_k||}{||\boldsymbol{x}_n - \mu_j||} \right)^{\frac{2}{m-1}}},$$

$$\mu_k = \frac{\sum_{n=1}^{N} r_{nk}^{m} \boldsymbol{x}_n}{\sum_{n=1}^{N} r_{nk}^{m}}.$$

  - K개 클러스터의 중심 간 거리 최대화
  - 클러스터 내 관측치와 클러스터 중심 간 거리 최소화
  - FCM은 관측치의 Membership 값을 {0,1} 사이의 실수로 보온.

- Deep Embedded Clustering , Improved Deep Embedded Clustering
  - Feature 공간인 X가 아닌, 오토인코더로 매핑된 Z로 클러스터링
  - IDEC는 디코더까지 추가하여 Z의 왜곡 방지

Max, Mean / I, LN, N, MM

- Max : Max pooling (Batch 중 Max값을 선택)
- Mean : Mean pooling (Batch 전체를 Mean하여 그 값을 선택)

- I : Identity normalization (정규화 안함)
- LN : Layer Normalization (Batch별 모든 feature를 정규화)
- N : standard Normalization (전체 분포를 평균 0 분산 1로)
- MM : Min-Max Normalization (최대 최소값을 0, 1로 정규화)

ACC, NMI, ARI

$$ACC = \frac{\sum_{i=1}^{n} \delta(\alpha_1, map(l_i))}{n},$$

- ACCuracy
  - ML에서 모델의 성능을 평가하기 위해 사용되는 대표적인 계산식
  - Dataset의 레이블(정답)과 비교하여, 전체 데이터 중 제대로 분류된 비율

$$ARI = \frac{\sum_{i,j}\binom{n_{ij}}{2} - \frac{\left[\sum_i \binom{n_{i\cdot}}{2} \sum_j \binom{n_{\cdot j}}{2}\right]}{\binom{n}{2}}}{\frac{1}{2}\left[\sum_i \binom{n_{i\cdot}}{2} + \sum_j \binom{n_{\cdot j}}{2}\right] - \frac{\left[\sum_i \binom{n_{i\cdot}}{2} \sum_j \binom{n_{\cdot j}}{2}\right]}{\binom{n}{2}}},$$

- Normalized Mutual Information
- Adjusted Rand Index
  - 두 계산식 모두 클러스터링 알고리즘의 성능을 평가하기 위해 사용.
  - 일반적으로 0(랜덤 군집화) ~ 1(최적), 1에 가까울수록 잘 분류된 결과임.

$$NMI(U,V) = \frac{MI(U,V)}{\sqrt{H(U)H(V)}},$$

실험

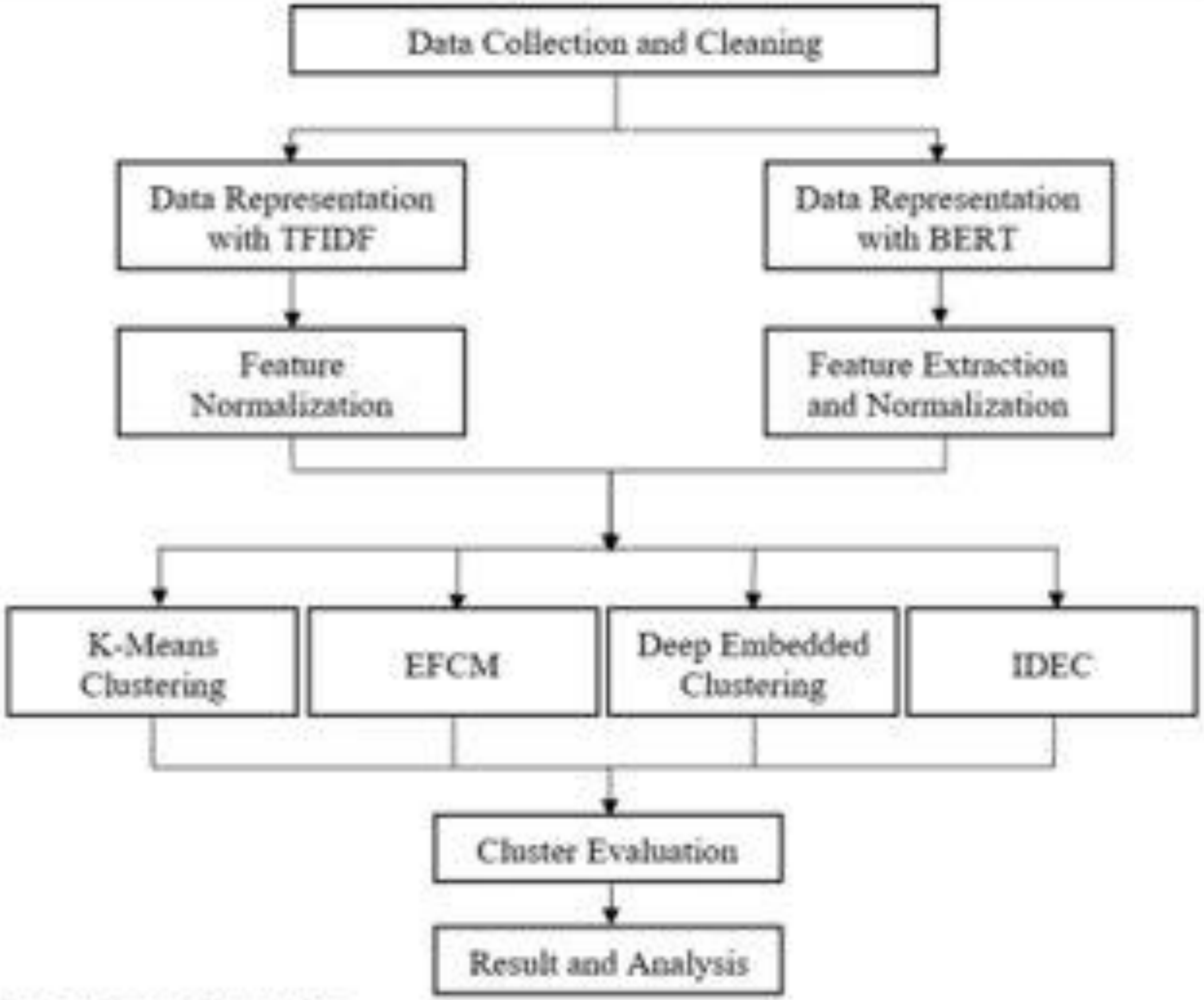

Data Collection and Cleaning
Data Representation with TFIDF
Data Representation with BERT
Feature Normalization
Feature Extraction and Normalization
K-Means Clustering
EFCM
Deep Embedded Clustering
IDEC
Cluster Evaluation
Result and Analysis

Fig. 1 Flowchart of Simulation

실험 (Dataset)

- **AGNews** : 뉴스 타이틀 및 내용으로 구성된 데이터, 뉴스 토픽으로 분류
  - 4개의 클래스로 분류되는 총 4000개의 데이터

- **Yahoo! Answers** : 질문 및 답변 데이터, 질문의 토픽으로 분류
  - 10개의 클래스로 분류되는 총 10000개의 데이터

- **Reuters** : 언론사에서 제공한 1987년도 뉴스 문서
  - 2개의 클래스로 분류되는 총 5859개의 데이터

실험 결과

- **AGNews**

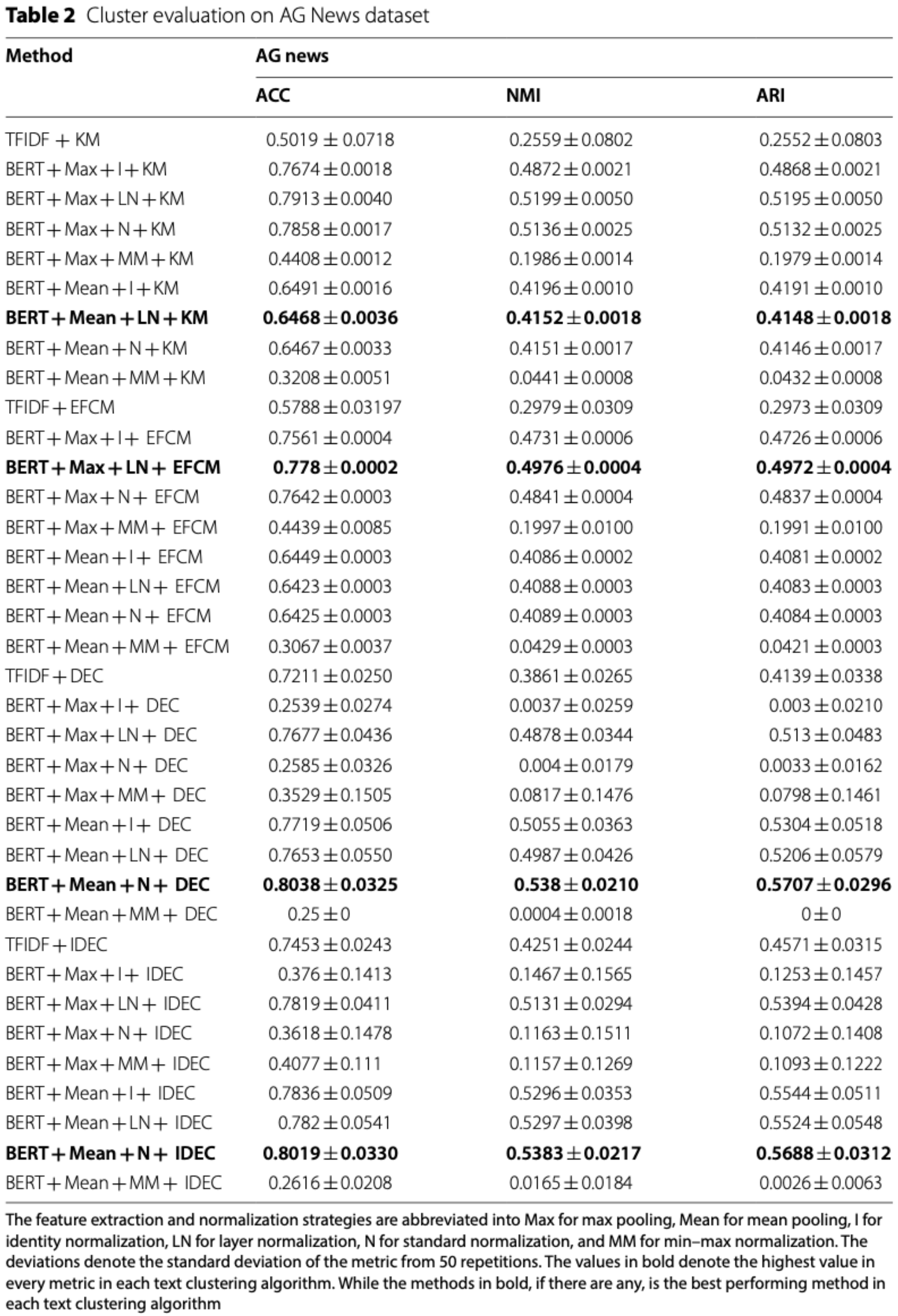


Table 2 Cluster evaluation on AG News dataset

Method	AG news		
	ACC	NMI	ARI
TFIDF + KM	0.5019 ± 0.0718	0.2559 ± 0.0802	0.2552 ± 0.0803
BERT + Max + I + KM	0.7674 ± 0.0018	0.4872 ± 0.0021	0.4868 ± 0.0021
BERT + Max + LN + KM	0.7913 ± 0.0040	0.5199 ± 0.0050	0.5195 ± 0.0050
BERT + Max + N + KM	0.7858 ± 0.0017	0.5136 ± 0.0025	0.5132 ± 0.0025
BERT + Max + MM + KM	0.4408 ± 0.0012	0.1986 ± 0.0014	0.1979 ± 0.0014
BERT + Mean + I + KM	0.6491 ± 0.0016	0.4196 ± 0.0010	0.4191 ± 0.0010
BERT + Mean + LN + KM	**0.6468 ± 0.0036**	**0.4152 ± 0.0018**	**0.4148 ± 0.0018**
BERT + Mean + N + KM	0.6467 ± 0.0033	0.4151 ± 0.0017	0.4146 ± 0.0017
BERT + Mean + MM + KM	0.3208 ± 0.0051	0.0441 ± 0.0008	0.0432 ± 0.0008
TFIDF + EFCM	0.5788 ± 0.03197	0.2979 ± 0.0309	0.2973 ± 0.0309
BERT + Max + I + EFCM	0.7561 ± 0.0004	0.4731 ± 0.0006	0.4726 ± 0.0006
BERT + Max + LN + EFCM	**0.778 ± 0.0002**	**0.4976 ± 0.0004**	**0.4972 ± 0.0004**
BERT + Max + N + EFCM	0.7642 ± 0.0003	0.4841 ± 0.0004	0.4837 ± 0.0004
BERT + Max + MM + EFCM	0.4439 ± 0.0085	0.1997 ± 0.0100	0.1991 ± 0.0100
BERT + Mean + I + EFCM	0.6449 ± 0.0003	0.4086 ± 0.0002	0.4081 ± 0.0002
BERT + Mean + LN + EFCM	0.6423 ± 0.0003	0.4088 ± 0.0003	0.4083 ± 0.0003
BERT + Mean + N + EFCM	0.6425 ± 0.0003	0.4089 ± 0.0003	0.4084 ± 0.0003
BERT + Mean + MM + EFCM	0.3067 ± 0.0037	0.0429 ± 0.0003	0.0421 ± 0.0003
TFIDF + DEC	0.7211 ± 0.0250	0.3861 ± 0.0265	0.4139 ± 0.0338
BERT + Max + I + DEC	0.2539 ± 0.0274	0.0037 ± 0.0259	0.003 ± 0.0210
BERT + Max + LN + DEC	0.7677 ± 0.0436	0.4878 ± 0.0344	0.513 ± 0.0483
BERT + Max + N + DEC	0.2585 ± 0.0326	0.004 ± 0.0179	0.0033 ± 0.0162
BERT + Max + MM + DEC	0.3529 ± 0.1505	0.0817 ± 0.1476	0.0798 ± 0.1461
BERT + Mean + I + DEC	0.7719 ± 0.0506	0.5055 ± 0.0363	0.5304 ± 0.0518
BERT + Mean + LN + DEC	0.7653 ± 0.0550	0.4987 ± 0.0426	0.5206 ± 0.0579
BERT + Mean + N + DEC	**0.8038 ± 0.0325**	**0.538 ± 0.0210**	**0.5707 ± 0.0296**
BERT + Mean + MM + DEC	0.25 ± 0	0.0004 ± 0.0018	0 ± 0
TFIDF + IDEC	0.7453 ± 0.0243	0.4251 ± 0.0244	0.4571 ± 0.0315
BERT + Max + I + IDEC	0.376 ± 0.1413	0.1467 ± 0.1565	0.1253 ± 0.1457
BERT + Max + LN + IDEC	0.7819 ± 0.0411	0.5131 ± 0.0294	0.5394 ± 0.0428
BERT + Max + N + IDEC	0.3618 ± 0.1478	0.1163 ± 0.1511	0.1072 ± 0.1408
BERT + Max + MM + IDEC	0.4077 ± 0.111	0.1157 ± 0.1269	0.1093 ± 0.1222
BERT + Mean + I + IDEC	0.7836 ± 0.0509	0.5296 ± 0.0353	0.5544 ± 0.0511
BERT + Mean + LN + IDEC	0.782 ± 0.0541	0.5297 ± 0.0398	0.5524 ± 0.0548
BERT + Mean + N + IDEC	**0.8019 ± 0.0330**	**0.5383 ± 0.0217**	**0.5688 ± 0.0312**
BERT + Mean + MM + IDEC	0.2616 ± 0.0208	0.0165 ± 0.0184	0.0026 ± 0.0063

The feature extraction and normalization strategies are abbreviated into Max for max pooling, Mean for mean pooling, I for identity normalization, LN for layer normalization, N for standard normalization, and MM for min–max normalization. The deviations denote the standard deviation of the metric from 50 repetitions. The values in bold denote the highest value in every metric in each text clustering algorithm. While the methods in bold, if there are any, is the best performing method in each text clustering algorithm

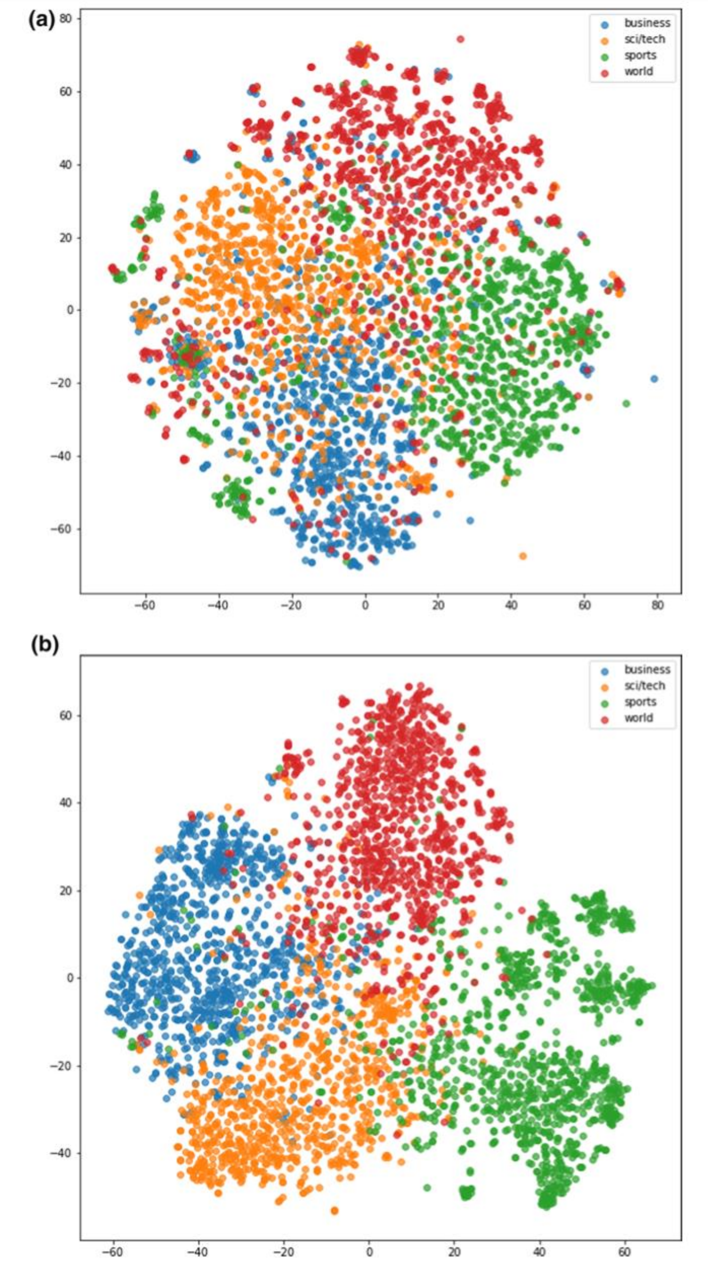


Fig. 3 t-SNE visualization of AG News ground truth label with text data representation from (**a**) TFIDF (**b**) BERT

실험 결과

- **Yahoo! Answers**

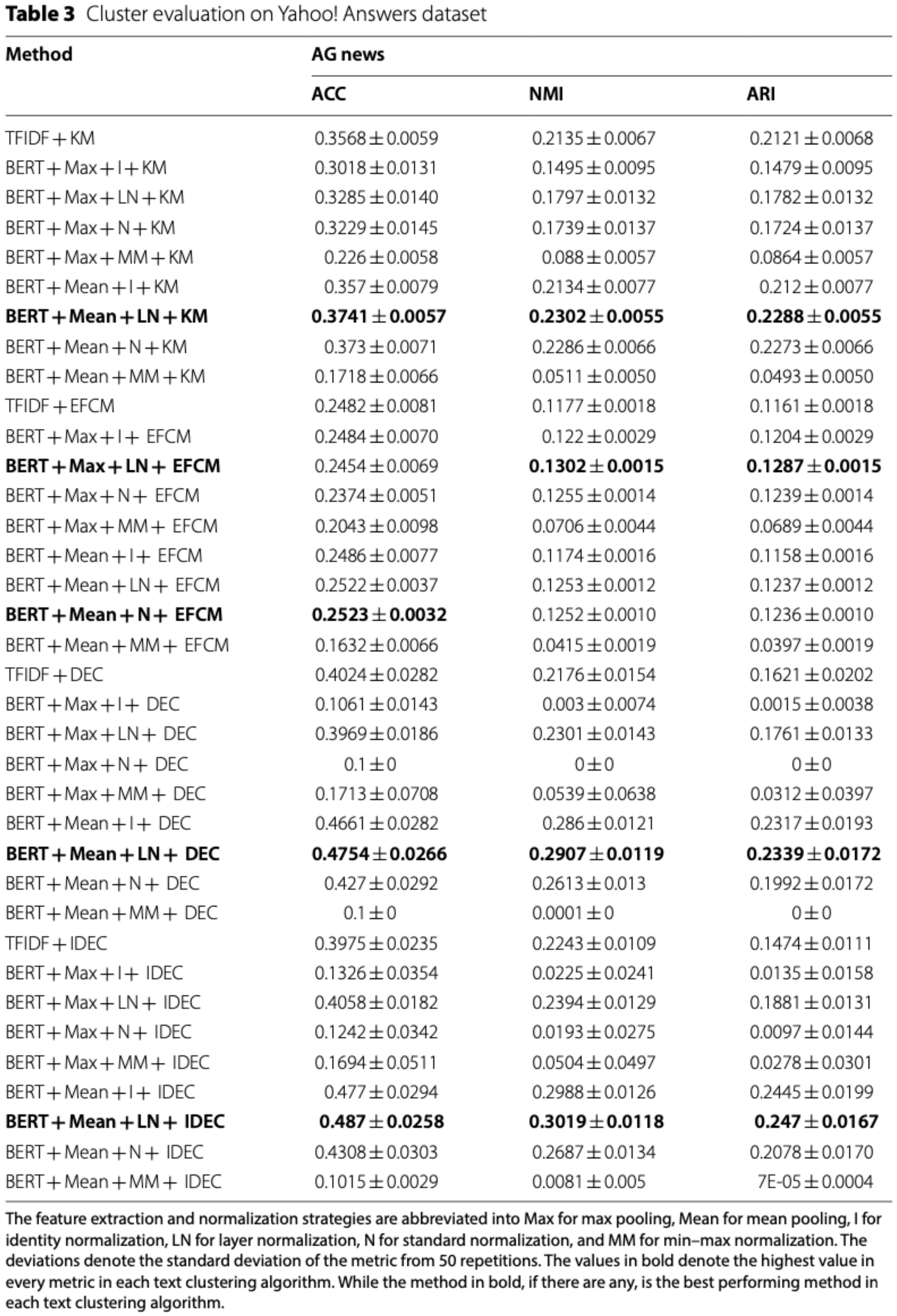


Table 3 Cluster evaluation on Yahoo! Answers dataset

Method	AG news		
	ACC	NMI	ARI
TFIDF + KM	0.3568 ± 0.0059	0.2135 ± 0.0067	0.2121 ± 0.0068
BERT + Max + I + KM	0.3018 ± 0.0131	0.1495 ± 0.0095	0.1479 ± 0.0095
BERT + Max + LN + KM	0.3285 ± 0.0140	0.1797 ± 0.0132	0.1782 ± 0.0132
BERT + Max + N + KM	0.3229 ± 0.0145	0.1739 ± 0.0137	0.1724 ± 0.0137
BERT + Max + MM + KM	0.226 ± 0.0058	0.088 ± 0.0057	0.0864 ± 0.0057
BERT + Mean + I + KM	0.357 ± 0.0079	0.2134 ± 0.0077	0.212 ± 0.0077
BERT + Mean + LN + KM	**0.3741 ± 0.0057**	**0.2302 ± 0.0055**	**0.2288 ± 0.0055**
BERT + Mean + N + KM	0.373 ± 0.0071	0.2286 ± 0.0066	0.2273 ± 0.0066
BERT + Mean + MM + KM	0.1718 ± 0.0066	0.0511 ± 0.0050	0.0493 ± 0.0050
TFIDF + EFCM	0.2482 ± 0.0081	0.1177 ± 0.0018	0.1161 ± 0.0018
BERT + Max + I + EFCM	0.2484 ± 0.0070	0.122 ± 0.0029	0.1204 ± 0.0029
BERT + Max + LN + EFCM	0.2454 ± 0.0069	**0.1302 ± 0.0015**	**0.1287 ± 0.0015**
BERT + Max + N + EFCM	0.2374 ± 0.0051	0.1255 ± 0.0014	0.1239 ± 0.0014
BERT + Max + MM + EFCM	0.2043 ± 0.0098	0.0706 ± 0.0044	0.0689 ± 0.0044
BERT + Mean + I + EFCM	0.2486 ± 0.0077	0.1174 ± 0.0016	0.1158 ± 0.0016
BERT + Mean + LN + EFCM	0.2522 ± 0.0037	0.1253 ± 0.0012	0.1237 ± 0.0012
BERT + Mean + N + EFCM	**0.2523 ± 0.0032**	0.1252 ± 0.0010	0.1236 ± 0.0010
BERT + Mean + MM + EFCM	0.1632 ± 0.0066	0.0415 ± 0.0019	0.0397 ± 0.0019
TFIDF + DEC	0.4024 ± 0.0282	0.2176 ± 0.0154	0.1621 ± 0.0202
BERT + Max + I + DEC	0.1061 ± 0.0143	0.003 ± 0.0074	0.0015 ± 0.0038
BERT + Max + LN + DEC	0.3969 ± 0.0186	0.2301 ± 0.0143	0.1761 ± 0.0133
BERT + Max + N + DEC	0.1 ± 0	0 ± 0	0 ± 0
BERT + Max + MM + DEC	0.1713 ± 0.0708	0.0539 ± 0.0638	0.0312 ± 0.0397
BERT + Mean + I + DEC	0.4661 ± 0.0282	0.286 ± 0.0121	0.2317 ± 0.0193
BERT + Mean + LN + DEC	**0.4754 ± 0.0266**	**0.2907 ± 0.0119**	**0.2339 ± 0.0172**
BERT + Mean + N + DEC	0.427 ± 0.0292	0.2613 ± 0.013	0.1992 ± 0.0172
BERT + Mean + MM + DEC	0.1 ± 0	0.0001 ± 0	0 ± 0
TFIDF + IDEC	0.3975 ± 0.0235	0.2243 ± 0.0109	0.1474 ± 0.0111
BERT + Max + I + IDEC	0.1326 ± 0.0354	0.0225 ± 0.0241	0.0135 ± 0.0158
BERT + Max + LN + IDEC	0.4058 ± 0.0182	0.2394 ± 0.0129	0.1881 ± 0.0131
BERT + Max + N + IDEC	0.1242 ± 0.0342	0.0193 ± 0.0275	0.0097 ± 0.0144
BERT + Max + MM + IDEC	0.1694 ± 0.0511	0.0504 ± 0.0497	0.0278 ± 0.0301
BERT + Mean + I + IDEC	0.477 ± 0.0294	0.2988 ± 0.0126	0.2445 ± 0.0199
BERT + Mean + LN + IDEC	**0.487 ± 0.0258**	**0.3019 ± 0.0118**	**0.247 ± 0.0167**
BERT + Mean + N + IDEC	0.4308 ± 0.0303	0.2687 ± 0.0134	0.2078 ± 0.0170
BERT + Mean + MM + IDEC	0.1015 ± 0.0029	0.0081 ± 0.005	7E-05 ± 0.0004

The feature extraction and normalization strategies are abbreviated into Max for max pooling, Mean for mean pooling, I for identity normalization, LN for layer normalization, N for standard normalization, and MM for min–max normalization. The deviations denote the standard deviation of the metric from 50 repetitions. The values in bold denote the highest value in every metric in each text clustering algorithm. While the method in bold, if there are any, is the best performing method in each text clustering algorithm.

실험 결과

- **Reuters**

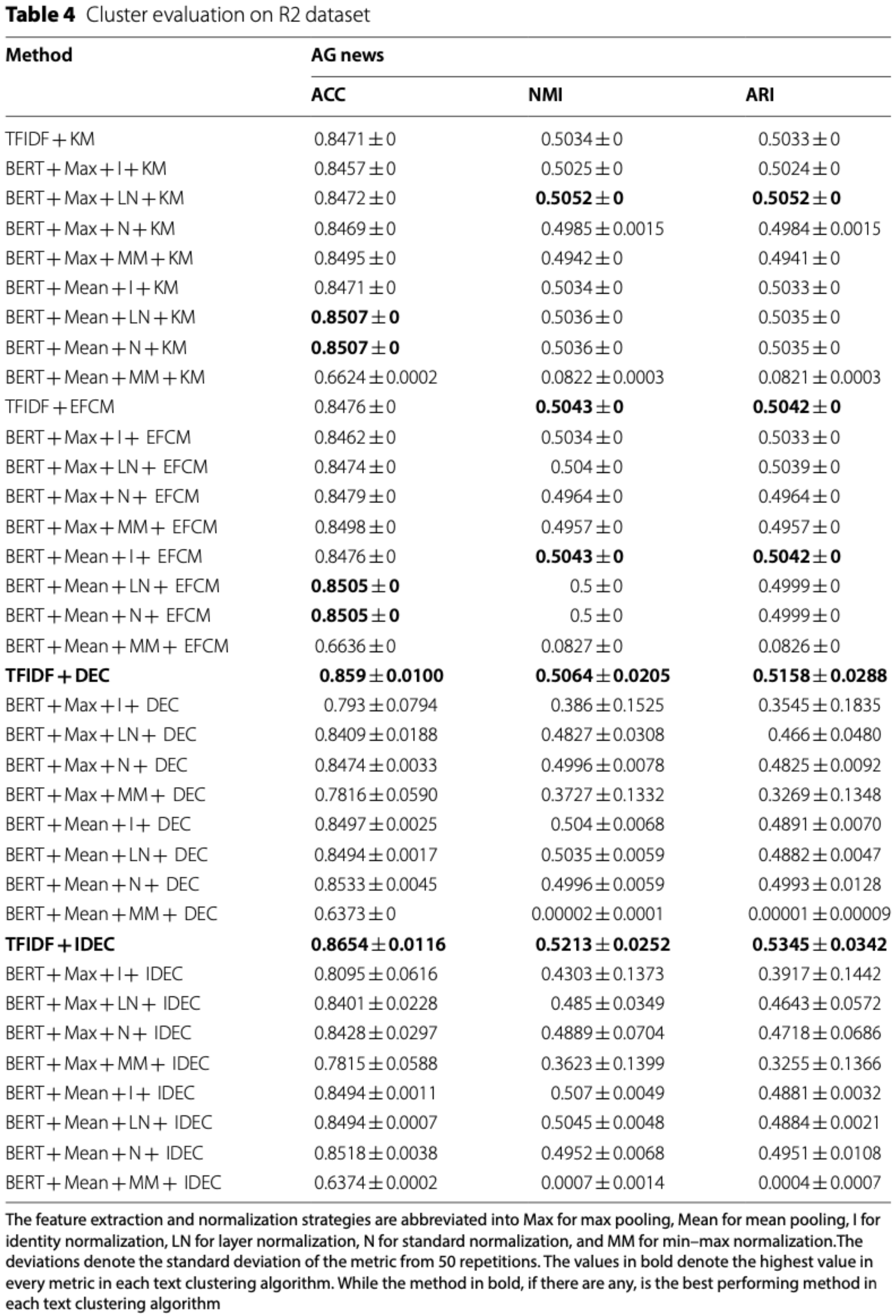


Table 4 Cluster evaluation on R2 dataset

Method	AG news		
	ACC	NMI	ARI
TFIDF + KM	0.8471 ± 0	0.5034 ± 0	0.5033 ± 0
BERT + Max + I + KM	0.8457 ± 0	0.5025 ± 0	0.5024 ± 0
BERT + Max + LN + KM	0.8472 ± 0	**0.5052 ± 0**	**0.5052 ± 0**
BERT + Max + N + KM	0.8469 ± 0	0.4985 ± 0.0015	0.4984 ± 0.0015
BERT + Max + MM + KM	0.8495 ± 0	0.4942 ± 0	0.4941 ± 0
BERT + Mean + I + KM	0.8471 ± 0	0.5034 ± 0	0.5033 ± 0
BERT + Mean + LN + KM	**0.8507 ± 0**	0.5036 ± 0	0.5035 ± 0
BERT + Mean + N + KM	**0.8507 ± 0**	0.5036 ± 0	0.5035 ± 0
BERT + Mean + MM + KM	0.6624 ± 0.0002	0.0822 ± 0.0003	0.0821 ± 0.0003
TFIDF + EFCM	0.8476 ± 0	**0.5043 ± 0**	**0.5042 ± 0**
BERT + Max + I + EFCM	0.8462 ± 0	0.5034 ± 0	0.5033 ± 0
BERT + Max + LN + EFCM	0.8474 ± 0	0.504 ± 0	0.5039 ± 0
BERT + Max + N + EFCM	0.8479 ± 0	0.4964 ± 0	0.4964 ± 0
BERT + Max + MM + EFCM	0.8498 ± 0	0.4957 ± 0	0.4957 ± 0
BERT + Mean + I + EFCM	0.8476 ± 0	**0.5043 ± 0**	**0.5042 ± 0**
BERT + Mean + LN + EFCM	**0.8505 ± 0**	0.5 ± 0	0.4999 ± 0
BERT + Mean + N + EFCM	**0.8505 ± 0**	0.5 ± 0	0.4999 ± 0
BERT + Mean + MM + EFCM	0.6636 ± 0	0.0827 ± 0	0.0826 ± 0
TFIDF + DEC	**0.859 ± 0.0100**	**0.5064 ± 0.0205**	**0.5158 ± 0.0288**
BERT + Max + I + DEC	0.793 ± 0.0794	0.386 ± 0.1525	0.3545 ± 0.1835
BERT + Max + LN + DEC	0.8409 ± 0.0188	0.4827 ± 0.0308	0.466 ± 0.0480
BERT + Max + N + DEC	0.8474 ± 0.0033	0.4996 ± 0.0078	0.4825 ± 0.0092
BERT + Max + MM + DEC	0.7816 ± 0.0590	0.3727 ± 0.1332	0.3269 ± 0.1348
BERT + Mean + I + DEC	0.8497 ± 0.0025	0.504 ± 0.0068	0.4891 ± 0.0070
BERT + Mean + LN + DEC	0.8494 ± 0.0017	0.5035 ± 0.0059	0.4882 ± 0.0047
BERT + Mean + N + DEC	0.8533 ± 0.0045	0.4996 ± 0.0059	0.4993 ± 0.0128
BERT + Mean + MM + DEC	0.6373 ± 0	0.00002 ± 0.0001	0.00001 ± 0.00009
TFIDF + IDEC	**0.8654 ± 0.0116**	**0.5213 ± 0.0252**	**0.5345 ± 0.0342**
BERT + Max + I + IDEC	0.8095 ± 0.0616	0.4303 ± 0.1373	0.3917 ± 0.1442
BERT + Max + LN + IDEC	0.8401 ± 0.0228	0.485 ± 0.0349	0.4643 ± 0.0572
BERT + Max + N + IDEC	0.8428 ± 0.0297	0.4889 ± 0.0704	0.4718 ± 0.0686
BERT + Max + MM + IDEC	0.7815 ± 0.0588	0.3623 ± 0.1399	0.3255 ± 0.1366
BERT + Mean + I + IDEC	0.8494 ± 0.0011	0.507 ± 0.0049	0.4881 ± 0.0032
BERT + Mean + LN + IDEC	0.8494 ± 0.0007	0.5045 ± 0.0048	0.4884 ± 0.0021
BERT + Mean + N + IDEC	0.8518 ± 0.0038	0.4952 ± 0.0068	0.4951 ± 0.0108
BERT + Mean + MM + IDEC	0.6374 ± 0.0002	0.0007 ± 0.0014	0.0004 ± 0.0007

The feature extraction and normalization strategies are abbreviated into Max for max pooling, Mean for mean pooling, I for identity normalization, LN for layer normalization, N for standard normalization, and MM for min–max normalization.The deviations denote the standard deviation of the metric from 50 repetitions. The values in bold denote the highest value in every metric in each text clustering algorithm. While the method in bold, if there are any, is the best performing method in each text clustering algorithm

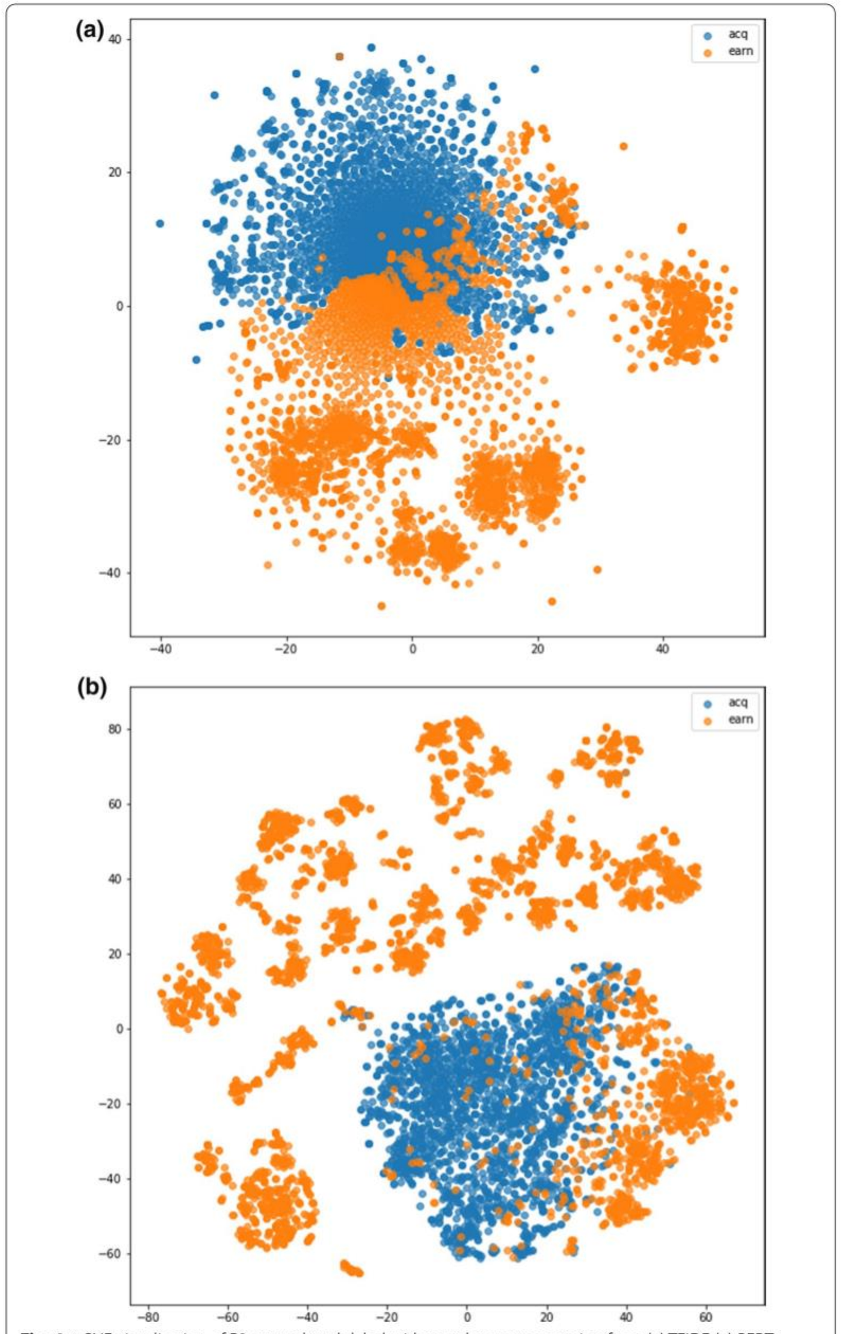


Fig. 4 t-SNE visualization of R2 ground truth label with text data representation from (**a**) TFIDF (**a**) BERT

결론

- 논문을 통해 성능이 좋은 알고리즘 조합에 대한 개략적인 정보와, NLP 비지도 학습의 실험 및 검증 방법에 대한 Insight를 얻을 수 있었음.
- 해당 실험에 사용된 Model, Dataset(docx, hwp 등에 embedding)을 참고하여 "포렌식을 위한 문서파일 분류기"를 실험할 예정.
- 추가로, BERT는 여전히 긴 문장의 경우 단어의 맥락(context) 파악이 힘들며, 이를 개선한 모델(CogLTX)을 제안하는 논문 존재. 해당 모델도 같이 실험 고려 (Dataset for "Classification": 20NewsGroups, Alibaba)
  - Ming Ding et al., "CogLTX: Applying BERT to Long Texts", NeurIPS, 2020