한국컴퓨터정보학회 동계학술대회 논문집 제29권 제1호 (2021. 1)

유튜브 메타정보를 이용한 자동 주제 분류 고찰

김용우*, 전성배O, 정유철*
O금오공과대학교 컴퓨터공학과,
*금오공과대학교 컴퓨터공학과
e-mail: {20121631, doqmflqo, jyc}@kumoh.ac.krO*

Analysis of Automatic Topic Classification using Youtube Meta Information

Yong-Woo Kim*, Seong-Bae JeonO, Yuchul Jung*
ODept. of Computer Engineering, Kumoh National Institute of Technology,
*Dept. of Computer Engineering, Kumoh National Institute of Technology

● 요 약 ●

Youtube 동영상 업로드 시, 사용자가 직접 주제를 설정해야 하는 어려움이 있다. 본 연구에서는 사용자가 입력하는 제목과 설명정보를 이용하여 자동으로 주제를 분류하는 연구를 진행하였다. 이를 위해 한국어기반의 컨텐츠 중 고빈도의 8개 주제 카테고리를 선정하고, 이를 1.3만건의 학습데이터를 크롤링을 통해 구축하였다. 또한, 다양한 알고리즘들에 대한 최대성능을 확인하기 위해 대표적인 텍스트 분류 방법인 SVM과 LSTM기법 및 BERT 모델기반 미세적용(fine-tuning)을 시도하였다. 결과적으로 Bert-multiligual (base)를 fine-tuning한 실험에서 최대 94%의 정확도를 확인하였다. 하지만, Youtube 동영상 특성상 여러 주제를 가진 것들이 상당수 존재하기에, 실제 체감정확도는 더 높을 것으로 기대된다.

키워드: 텍스트 분류(text classification), 버트(BERT)

I. Introduction

현재 유튜브에 영상을 업로드 할 때 영상 게시자가 영상의 카테고리를 수동으로 설정을 해주어야 한다.

이를 신경쓰지 않고 영상을 업로드하면 카테고리가 잘못 설정되어 다른 영상에 비해 일반 사용자들에게 노출도가 낮아질 수도 있다. 이를 방지하기 위해, 유튜브 카테고리 자동 분류를 위해 제목(title)과 설명(description)등의 메타정보를 활용한 자동 주제 분류를 시도하였으며, BERT 미세조정기법 중 BERT-multilingual (base)를 기반으로 한 모델이 가장 우수한 성능을 보였다.

II. Preliminaries

텍스트 분류(Text Classification)는 텍스트를 입력으로 받아, 텍스트가 어떤 종류의 범주에 속하는지를 구분하는 작업을 말한다. 관련된 기술로는 나이브 베이즈(Naive Bayes) [1], Support Vector Machines (SVM) [2], LSTM 등이 있고, Bert 모델을 finetuning하여 분류하는 방법이 있다.

나이브 베이즈는 A가 일어날 확률을 $P(A)$, B가 일어날 확률을 $P(B)$, A가 일어나고나서 B가 일어날 확률을 $P(B|A)$, B가 일어나고나서 A가 일어날 확률을 $P(A|B)$라고 할 때, $P(B|A)$를 쉽게 구할 수 있는 상황이라면, 아래와 같은 식(1)을 통해 $P(A|B)$를 구할 수 있다.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (1)$$

나이브 베이즈는 위의 식을 사용하여 텍스트를 분류한다.

RNN은 관련 정보와 그 정보를 사용하는 지점 사이 거리가 멀 경우 역전파시 그래디언트가 점차 줄어 학습능력이 크게 저하되는 vanishing gradient problem을 발생시키는데, 이를 극복하기 위해 LSTM [3]이 고안되었다.

SVM은 새로운 점이 나타나면 경계의 어느 쪽에 속하는지 확인해서 분류 과제를 수행한다. 이것을 분류하는 선을 결정 경계라고 하는데, 2차원에서 3차원으로 가면 결정 경계는 선에서 면이 된다. 이를 반복해 고차원의 결정 경계인 초평면을 그어 분류 과제를 수행한다.

Bidirectional Encoder Representations from Transformer (BERT) [4] 모델은 이미 학습된 모델을 기반으로 새로운 task의 성능을 더 올려주는 훈련된 언어 모델(Pre-trained Language Model)이다. 대표적으로 Google에서 배포한 Multilingual Model [5]이 있으며, 한국어 모델로는 KoBERT [6], HanBERT, KoELECTRA [7] 등이 있다.

III. The Proposed Scheme

1. Suggestion Method

1.1 Dataset 구축

Youtube상의 존재하는 text 데이터는 제목(title)과 설명(description) 그리고 label에 해당하는 주제가 있다. Youtube는 자체적으로 15종류(영화/애니메이션, 자동차/교통, 음악, 애완동물/동물, 스포츠, 여행/이벤트, 게임, 인물/블로그, 코미디, 엔터테인먼트, 뉴스/정치, 노하우/스타일, 교육, 과학기술, 비영리/사회운동)의 주제(topic)를 가지고 있다. 하지만 주제(topic)의 모호성과 데이터의 불균형 문제로 총 8종류(자동차/교통, 음악, 애완동물/동물, 스포츠, 게임, 엔터테인먼트, 뉴스/정치, 교육)의 주제(topic)를 선정하였다.

우리는 무작위 Youtube 데이터를 Selenium을 이용해서 Crawling하여 각 주제(topic)당 약 5,000개의 데이터를 확보 하였고 통계는 (Fig 1)과 (Table 1)과 같다.

Table 1. Statistics of Our Youtube Dataset

Num. of Topic Categories	Average Length (# of words)	Max Length (# of words)	Train (# of docs)	Test (# of docs)
8	147	1523	30964	10322

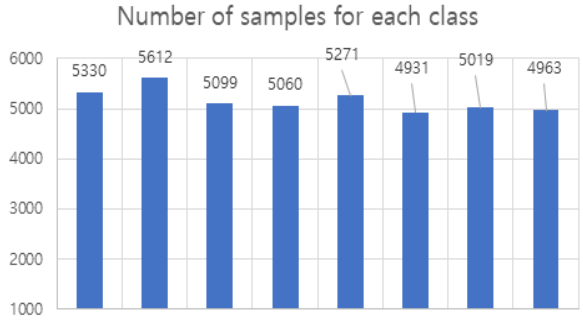

Fig. 1. Numers of Docs. in Each Topic Category

1.2 사용한 분류기법들

1.2.1 고전적 classifier 기법

Scikit-learn의 SVM, NB와 LSTM 방법을 이용해서 우리가 구축한 dataset을 학습시켜 정확도를 확인하고, 비교한다.

1.2.2 BERT기반 fine-tuning

다양한 pre-trained된 BERT model을 기반으로 우리 dataset을 finetuning하여 model을 생성하고 그 성능을 확인한다.

2. Experiment result & Analysis

2.1 Classic classifier techniques

2.1.1 Navie Bayes(NB)

	precision	recall	f1-score	support
Education	0.92	0.82	0.87	732
Entertainment	0.90	0.88	0.89	789
Game	0.92	0.84	0.87	722
Music	0.79	0.82	0.81	645
News	0.90	0.92	0.91	750
Pets	0.86	0.95	0.90	745
Sports	0.96	0.94	0.95	759
Vehicle	0.90	0.97	0.94	773
accuracy			0.89	5915
macro avg	0.89	0.89	0.89	5915
weighted avg	0.90	0.89	0.89	5915

Fig. 2. Results of NB Algorithm

NB는 모든 사건들이 독립이 아니어도 독립으로 가정하여 확률을 계산하는 모델이다. 이러한 한계 때문에 가장 낮은 정확도가 나온 것 같다.

2.1.2 Support Vector Machines (SVM)

	precision	recall	f1-score	support
Education	0.92	0.88	0.90	732
Entertainment	0.93	0.88	0.91	789
Game	0.93	0.89	0.91	722
Music	0.84	0.90	0.87	645
News	0.94	0.94	0.94	750
Pets	0.90	0.97	0.93	745
Sports	0.96	0.96	0.96	759
Vehicle	0.97	0.98	0.97	773
accuracy			0.93	5915
macro avg	0.92	0.92	0.92	5915
weighted avg	0.93	0.93	0.93	5915

Fig. 3. Results of SVM Algorithm

SVM은 적은 데이터로도 높은 정확도를 보여주는 모델로 잘 알려져 있다. 우리의 dataset도 많은 양의 정보를 가지지 않았기 때문에 다른 모델보다 SVM이 상대적으로 높은 정확도를 보여준 것으로 생각된다.

2.1.3 Long-short Term Memory (LSTM)

	precision	recall	f1-score	support
Education	0.87	0.82	0.84	707
Entertainment	0.88	0.88	0.88	839
Game	0.86	0.90	0.88	678
Music	0.87	0.81	0.84	647
News	0.84	0.93	0.89	748
Pets	0.92	0.91	0.91	750
Sports	0.94	0.92	0.93	736
Vehicle	0.97	0.97	0.97	810
accuracy			0.89	5915
macro avg	0.89	0.89	0.89	5915
weighted avg	0.90	0.89	0.89	5915

Fig 4. LSTM 분석 결과

LSTM은 거시적인 과거 데이터를 고려하여 미래 데이터를 예측하는데 용이한데 우리의 dataset은 데이터가 충분하지 않았기 때문에 BERT pre-trained model을 사용한 것보다 정확도가 낮게 나온 것이라 생각된다.

2.2 BERT finetuning techniques

Table 2. Experimental Results of BERT Fine-tuning

Models	Learning Rate	Vocab. Size	Batch Size	Accuracy
KoBERT	5e-05	8,002	32	0.8923
Bert-base-Multilingual	5e-05	119,547	32	0.9400
KoELECTRA-base-v3	5e-05	35,000	16	0.9388
KoELECTRA-small-v3	5e-05	35,000	16	0.9232

SKTBrain의 KoBERT model로 학습했을 경우 정확도는 고전적인 기법과 크게 차이가 없는 0.8923을 보였는데 이는 vocab size가 8,002개로 가장 낮았기 때문에 이런 결과가 나온 것으로 예상 된다.

Google의 Multilingual model은 여러 개의 언어를 학습한 모델이고, 우리가 구축한 Youtube dataset은 영어를 포함한 dataset이기 때문에 0.9400으로 가장 높은 성능을 보였다.

KoELECTRA는 base와 small 두 가지 모델로 학습을 진행시켰는데, small 모델은 base모델에 비해 Embedding Size와 Hidden Size, heads의 수치를 경량화 한 모델이다. KoELECTRA는 한국어 Vocabulary data가 35,000개로 충분한 pre-trained이 되어 있어, base 0.9388 small 0.9232의 정확도를 보여주었다.

IV. Conclusions

고전적인 classifier 기법들 보다 Pre-trained 된 BERT model을 사용했을 경우가 일반적으로 정확도가 높은 것을 확인하였다.

유튜브는 15개의 주제로 나눠져 있긴 하지만 (Table 3)과 같이 여러 개의 주제를 공통으로 가진 영상도 있기 때문에 임의의 100개의 영상을 test해 보았을 때 95개의 정답과 5개의 오답이 있었는데, 확인해 본 결과 위 오답 5개 중 3개는 해당 영상이 여러 개의 주제(topic)를 가진 영상이었다. 그렇기 때문에 실제로 적용했을 때 실험 결과로 나온 Accuracy보다 좋은 결과를 얻을 수 있을 것으로 예상된다.

Table 3. Video containing various topics

영상 제목	Topic	KoBERT model predict
8부 팀과 토트넘 대결…"손흥민을 상대한다고?" (2020.12.01/뉴스데스크/MBC)	Sports	News
여자 컬링 준결승 한일전 주요장면.. 1엔드부터 11엔드까지 (하이라이트)	Sports	News

REFERENCES

[1] Scikit-learn - Naive Bayes [Online]. Available: https://scikit-learn.org/stable/modules/naive_bayes.html

[2] Scikit-learn – Support Vector Machines [Online]. Available: https://scikit-learn.org/stable/modules/svm.html

[3] LSTM - [Online]. Available: https://ratsgo.github.io/natural%20language%20processing/2017/03/09/rnnlstm/

[4] BERT - [Online]. Available: https://ebbnflow.tistory.com/151

[5] An Empirical Analysis of Naver Movie Review with Hugging Face BERT - [Online]. Available: https://colab.research.google.com/drive/1tIf0Ugdqg4qT7gcxia3tL7und64Rv1dP?hl=ko

[6] KoBERT - [Online]. Available: https://github.com/SKTBrain/KoBERT

[7] KoELECTRA - [Online]. Available: https://github.com/monologg/KoELECTRA