다중클래스 한국어 감성분석에서 클래스 불균형과 손실 스파이크 문제 해결을 위한 기법

박제윤[1○*], 양기수[1*], 박예원[2*], 이문기[3*], 이상원[4*], 임수연[5*], 조재훈[6*], 임희석[1*]
[1]고려대학교, [2]숙명여자대학교, [3](주)대륜이엔에스, [4]인천대학교, [5]한국과학기술연구원, [6]연세대학교
{k4ke,willow4,limhseok}@korea.ac.kr, yw.park0503@gmail.com
mglee@daeryunens.co.kr, leo503801@inu.ac.kr, yeen0707@kist.re.kr, jaycho0309@yonsei.ac.kr

Methods For Resolving Challenges In Multi-class Korean Sentiment Analysis

Jeiyoon Park[1○*], Kisu Yang[1*], Yewon Park[2*], Moongi Lee[3*], Sangwon Lee[4*], Sooyeon Lim[5*], Jaehoon Cho[6*], Heuiseok Lim[1*]
[1]Korea University, [2]Sookmyung Women's University, [3]Daeryun E&S, [4]Incheon University, [5]KIST, [6]Yonsei University

요 약

오픈 도메인 대화에서 텍스트에 나타난 태도나 성향과 같은 화자의 주관적인 감정정보를 분석하는 것은 사용자들에게서 풍부한 응답을 이끌어 내고 동시에 제공하는 목적으로 사용될 수 있다. 하지만 한국어 감성분석에서 기존의 대부분의 연구들은 긍정과 부정 두개의 클래스 분류만을 다루고 있고 이는 현실 화자의 감정 정보를 정확하게 분석하기에는 어려움이 있다. 또한 최근에 오픈한 다중클래스로된 한국어 대화 감성분석 데이터셋은 중립 클래스가 전체 데이터셋의 절반을 차지하고 일부 클래스는 사용하기에 매우 적은, 다시 말해 클래스 간의 데이터 불균형 문제가 있어 다루기 굉장히 까다롭다. 이 논문에서 우리는 일곱개의 클래스가 존재하는 한국어 대화에서 세션들을 효율적으로 분류하는 기법들에 대해 논의한다. 우리는 극심한 클래스 불균형에도 불구하고 76.56 micro F1을 기록하였다.

주제어: 감성분석, 한국어 감성분석, 다중클래스 한국어 감성분석

1. 서론

감성분석(Sentiment analysis)이란 텍스트에서 선호도나 상태등의 주관적인 감정 정보를 추측하고 분류하는 연구분야이다. 감성분석은 상품평 분석을 통한 고객의 니즈분석 [1], 광고 효과 분석 [2], 영화평 분석 [3], 그리고 풍부한 대화 시스템 응답 [4] 등 다양한 분야에 적용될 수 있다. 특히 최근 문자만을 통한 비대면 소통이 늘면서 비언어적인 요소가 부재한 환경에서 화자간의 명확한 의미 전달에 도움이 될 수 있다 [5].

한국어 감성분석은 어순이 자유로워 구문의 구조와 의미를 분석하는 작업이 복잡하고, 대화문에서는 주어나 명사의 생략이 많아 모호한 문장이 많으며, 한 문장이 다양한 의미를 가지는 경우가 매우 많다. 최근 한국어로 된 large-scale 모델들의 발전으로 한국어에서도 이러한 문제점들을 다룬 감성분석 관련 연구가 활발하게 진행되고 있다 [6, 7, 8].

하지만, 대부분의 연구는 실제 사용자의 감정을 긍정과 부정 두 개의 클래스 만으로 분류하여 연구하고 있다 [9, 10]. 언어에 내포된 다양한 감정 정보를 긍정과 부정 두가지 종류로 분류하는 것으로는 현실 화자의 감정 정보를 정확하게 분석하기에 어려움이 있다.

이러한 문제점을 해결하기 위해 최근에 AI Hub에서 다중클래스 모델 학습을 위한 한국어 감정 정보가 포함된 연속적 대화 데이터 셋과 단발성 대화 데이터 셋을 공개하였다. 그림 1은 연속적 대화 데이터셋의 일부분이다. 데이터셋은 중립, 혐오, 공포, 분노, 놀람, 슬픔, 그리고 기쁨 (또는 행복) 총 일곱개의

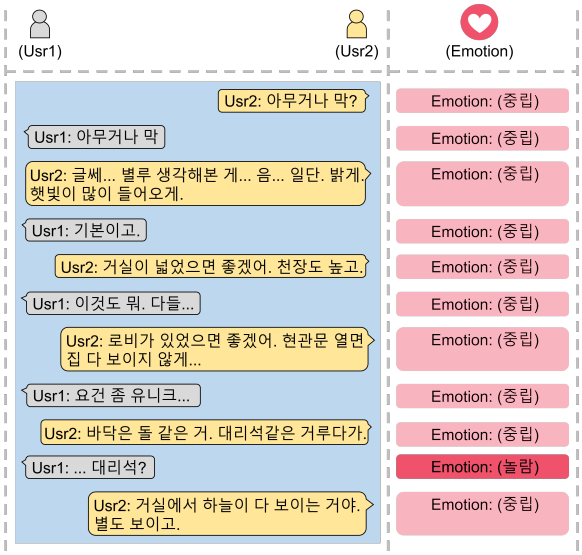


그림 1. AI Hub 한국어 감정정보가 포함된 연속적 대화 세션 예시.

클래스로 이루어져 있다. 하지만, 그림 1과 같이 화자간의 대화 특성상 대부분의 데이터는 중립 정보를 가지게 되고 이는 전체 데이터에서 클래스 불균형 문제를 발생시킨다. 이는 가장 많은 비중을 차지하는 중립보다 훨씬 적은 클래스 수를 가지지만 다른 감정 클래스가 더욱 중요한 다중클래스 감성분석 문제에서 모델 학습시 중립 클래스에 지나치게 편향되는 문제를 야기한다. 또한, 학습과정에서 학습 데이터의 label이 정렬된 형태로 순차적이게 들어가게 되고 이는 손실스파이크(Loss spikes)문제를 발생시켜 학습이 진행되지 않는 문제를 일으킨다.

*These authors contributed equally.

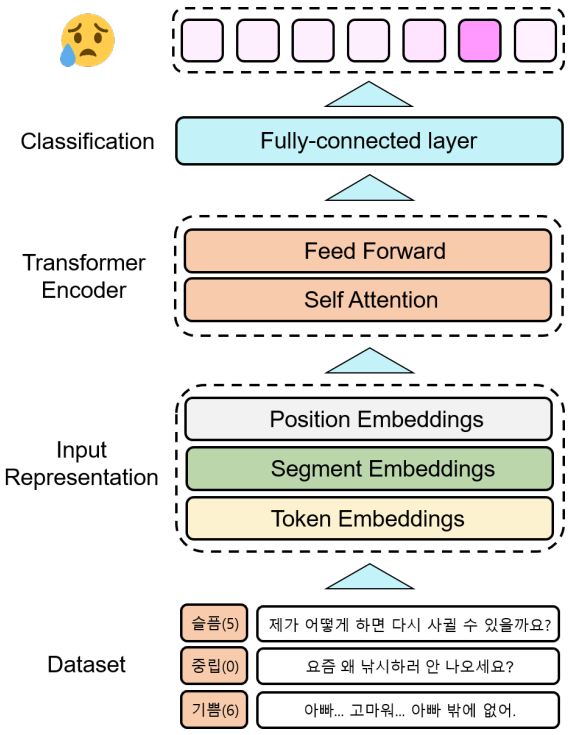


그림 2. 다중클래스 감성분석 모델의 개요.

이 논문에서 우리는 기존 연구들에서 많이 다루어지지 않은 한국어 다중클래스 감성분석과 여기에서 발생하는 문제점들을 해결하기 위한 방안들에 대해 논의한다. 학습 과정에서 데이터 클래스의 불균형 문제로 인한 학습이 원활하게 되지 않는 문제를 해결하기 위해 기존의 Cross entropy에 클래스 수만큼 패널티를 주어서 학습하였다. 또한 학습과정에서 연속적 대화 데이터셋과 단발성 대화 데이터셋을 합치고 무작위로 섞어서 데이터가 레이블에 순차적으로 학습되는 문제를 해결하였다. 우리는 극심한 클래스 불균형에도 불구하고 76.56 micro F1을 기록하였다.

2. BERT

자연어를 처리하고 가공하기 위해서는 컴퓨터가 이해할 수 있도록 단어를 벡터로 나타내는 단어 임베딩이 필수적이다. 이러한 단어 임베딩을 구현하기 위한 방법으로는 최근 Word2Vec [11], GloVe [12], ELMo [13], BERT [14] 등 다양한 방법론이 제시되고 있다.

그 중 가장 최근에 발표된 BERT는 일부 성능 평가에서 인간보다 더 높은 정확도를 보일 정도로 발전된 모델이다. BERT는 그림 2 과 같이 세 단계 과정을 거친다. (1) Input : 기존 단어를 쪼갠 단어조각을 단위로 임베딩하는 Token Embedding과 두 문장을 학습하는 Sentence Embedding, 그리고 Transformer의 Position Embedding 이 세가지의 합으로 Input을 구성한다. (2) Pre-Training 단어 중 일부를 [MASK] Token으로 바꾼 후 이 [MASK]를 predict하는 'Masked LM'과 문장의 관계를 predict하는 'Next Sentence prediction'을 수행한다. (3) Fine-Tuning, 즉 실제 자연어처리 문제를 풀며 추가 모델을 쌓아 모델에 조정을 가한다.

BERT는 다른 모델에 비해 우수한 언어 이해 능력을 가진 덕분에 최근 많은 자연어처리 연구에 적용되고 있고 특히 감성분석에도 활발하게 적용되고 있다 [15]. 우리는 이러한 점에서 영감을 받아 SKT에서 공개한 한국어에 적용 가능한 BERT를 활용한다[1].

감정 클래스	데이터 갯수
중립	48631
혐오	5650
공포	5566
분노	9299
놀람	10766
슬픔	7241
기쁨 (행복)	7068
계	94221

표 1. AI Hub 한국어 감정정보가 포함된 대화 데이터셋의 각 감정 클래스당 데이터 수.

3. 실험

3.1 다중클래스 한국어 감성분석 데이터

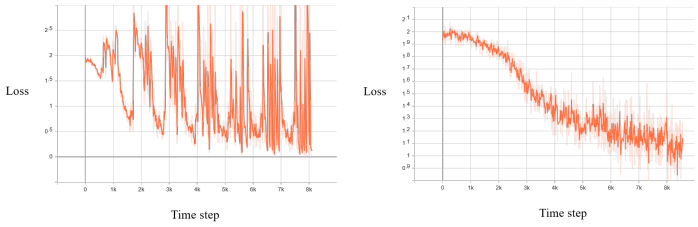
이 논문에서 우리는 AI-hub에서 공개한 다중클래스 한국어 감정 정보가 포함된 연속적인 대화 데이터셋과 단발성 대화 데이터셋을 사용하였다[2]. 우리는 학습과정에서 이 데이터셋을 표 1와 같이 하나로 합쳐서 사용하였다. 표 1에서 볼 수 있는 것처럼, 전체 9만여 대화 세션들 중 중립이 절반 이상을 차지한다. 이는 모델 학습시 중립 클래스에 지나치게 편향되는 불균형 문제를 발생 시킨다. 이는 한 클래스에 속하는 데이터 수가 적지만 중요한 감정 정보를 가지는 다중클래스 감성분석 모델에 치명적인 문제를 발생시킨다.

3.2 클래스 불균형

기존 한국어 감성분석에서는 주로 긍정과 부정으로 나누는 문제를 해결하기 위해 Cross entropy loss function을 사용하였다. Cross entropy loss function의 식은 다음과 같다:

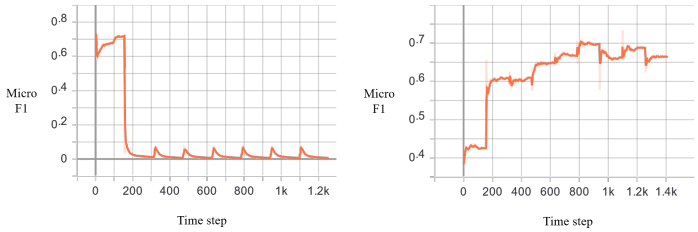
$$H_p(y) = -\sum_{i}^{N} y(x_i) \log p(x_i) \quad (1)$$

[1]https://github.com/SKTBrain/KoBERT
[2]https://aihub.or.kr



(a) (b)

그림 3. (a) 손실 스파이크가 발생했을 때의 손실함수, (b) 손실 스파이크를 해결했을 때의 손실함수.



(a) (b)

그림 4. (a) 손실 스파이크가 발생했을때의 성능, (b) 손실 스파이크를 해결했을때의 성능.

여기서 N은 전체 데이터 수, x_i는 각각의 데이터, 그리고 y는 데이터의 클래스를 의미한다. 하지만, 다중클래스 감성분석에서 Equation 1을 사용할 경우 클래스 불균형에 따른 잘못된 학습을 할 가능성이 매우 크다. 다시말해, 중립 클래스에 편향된 모델 파라미터를 학습할 가능성이 크다. 한국어 다중클래스 감성분석에서 중립 클래스의 데이터가 가장 많은 비중을 차지하지만 실제로 더 많은 정보를 담고 있는 것은 중립 외의 데이터 수가 적은 감정 정보 이므로 클래스 불균형 문제를 해결하기 위한 기법은 필수적이다.

우리는 Weighted cross entropy (WCE)를 통해 다중클래스 한국어 감성분석에서 대화의 특성상 존재할 수 밖에 없는 클래스의 불균형 문제를 해결하였다. 즉, 우리는 WCE를 이용하여 아래와 같이 클래스 수에 따라 패널티를 주었다:

$$\tilde{H}_p(y_k) = -\frac{N}{A(k)} \sum_{i}^{n_k} y_k(x_i) \log p(x_i) \tag{2}$$

$$\text{where} \quad N = \sum_j x_j \quad \text{and} \quad A(k) = \sum y_k \tag{3}$$

여기서 k는 각 클래스에 대응되는 수 (i.e. 0 to 6), y_k는 감정 클래스, 그리고 n_k는 각 클래스에 포함된 데이터 수를 의미한다. 우리는 WCE 손실함수를 통해 클래스가 적은 편에 속하는 감정정보 데이터들도 잘 학습시켜 편향되지 않은 다중클래스 한국어 감성분석 모델을 구축하였다.

Method	Micro F1
KoBERT + Unshuffled	0.008
KoBERT + Shuffled	71.560
Ours	**76.560**

표 2. 경험적인 실험 결과.

3.3 손실 스파이크

손실 스파이크 (Loss spikes)란 학습 과정에서 손실 함수의 값이 매 epoch마다 크게 튀는 현상을 말한다. 손실 스파이크는 원래 Adam optimizer에서 mini-batch stochastic gradient descent를 사용할때 어느정도는 피할 수 없는 현상이지만, AI-hub에서 오픈한 한국어 다중클래스 데이터를 사용할 경우 그림 3a에서와 같이 손실함수 값이 지속적으로 매우 크게 튀는 것을 볼 수가 있다. 손실 스파이크가 발생할 경우 그림 4a처럼 학습이 아예 진행되지않는 문제가 발생한다. 우리는 이러한 문제를 해결하기 위해 두가지 데이터셋을 합치고 발화문장과 레이블 쌍을 메모리에 넣고 랜덤으로 섞어 일정 배치만큼 가져와서 학습시키는 방법을 사용하였다. 그 결과 그림 3b처럼 손실함수가 감소하는 것을 확인할 수 있고, 그림 4b처럼 학습이 잘 진행되는 것을 확인할 수 있다.

4. 결과

표 2 AI-hub 다중클래스 대화 데이터셋에서의 실험 결과를 나타낸다. 또한 우리는 클래스 불균형에 대한 정확한 측정을 위해 F1 score에 클래스 불균형 정도가 반영된 Micro F1을 측정 지표로 사용하였다. 손실 스파이크를 해결하지 않은 모델의 경우 그림 4a와 표 2에서 볼 수 있듯이 아예 학습이 진행되지 못하는 것을 볼 수있다. 손실 스파이크를 해결한 모델에 대해서는 어느정도 잘 학습이 되었으나 손실 함수에 클래스 가중치를 주어 학습을 할 경우 더 높은 성능을 나타내는 것을 확인할 수 있다.

5. 결론

본 논문에서는 기존 긍정과 부정, 이 두 클래스 분류만을 다루었던 한국어 감성분석에 대해 BERT 모델을 활용하여 일곱개의 다양한 클래스로 분류하는 방법을 제안하였다. 이 기법에서는 한국어 대화 데이터셋의 클래스 간 데이터 불균형 문제를 해결하기 위하여, Shuffle과 Weighted Cross Entropy를 활용함으로서 편중된 데이터셋으로 인해 발생되는 문제들을 감소시키고 성능을 향상시킬 수 있었다.

우리는 더 나아가 그림 5과 같이 이 모델을 통해 한국어 대화에서의 감성 분석의 적용 사례로써 사용자의 감정에 적합한 이모지를 자동으로 추천되는 시스템을 간단하게 구현해보았

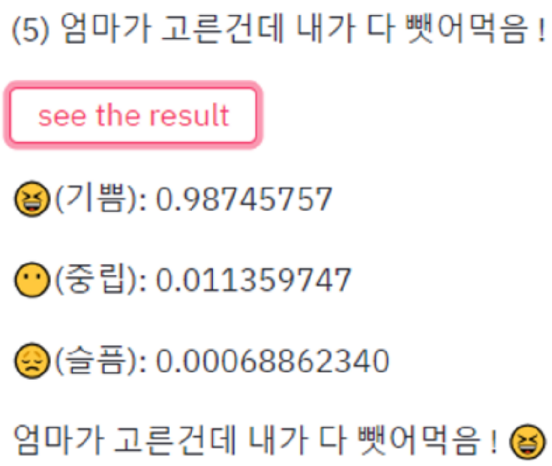


그림 5. 다중클래스 감성분석 활용 예시

다. 향후 연구로는 한국어에 적합한 이모지 추천을 통해 문자 위주의 비대면 소통시 비언어적 요소가 없는 환경을 어느정도 완화시킬 수 있을거라고 기대한다.

감사의 글

"본 연구는 과학기술정보통신부 및 정보통신기술기획평가원의 대학ICT연구센터지원사업의 연구결과로 수행되었음" (IITP-2020-2018-0-01405). "이 논문은 2017년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No.NRF-2017M3C4A7068189)."

참고문헌

[1] 박현정 외, "CNN을 적용한 한국어 상품평 감성분석," *지능정보연구*, Vol. 24, pp. 59–83, 2018.

[2] 김세진 외, "인터넷 용어의 감성 분석을 통한 동영상 광고 효과 분석 시스템 설계," *정보과학회논문지*, Vol. 46, pp. 919–925, 2019. [Online]. Available: http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE08769609

[3] 박천음 외, "BERT 기반 Variational Inference와 RNN을 이용한 한국어 영화평 감성 분석," *정보과학회 컴퓨팅의 실제 논문지*, Vol. 25, pp. 552–558, 2019. [Online]. Available: http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE09233106

[4] A. Rinaldi, O. Oseguera, J. Tuazon, and A. C. Cruz, "End-to-end dialogue with sentiment analysis features," *HCI International 2017 – Posters' Extended Abstracts*, C. Stephanidis, Ed., pp. 480–487, 2017.

[5] J. O'Neill and D. Martin, "Text chat in action," *Proceedings of the 2003 international ACM SIGGROUP conference on Supporting group work*, pp. 40–49, 2003.

[6] 민진우 외, "RoBERTa를 이용한 한국어 자연어처리: 개체명 인식, 감성분석, 의존파싱," *한국정보과학회 학술발표논문집*, Vol. , pp. 407–409, 2019. [Online]. Available: http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE09301603

[7] 박광현 외, "BERT를 이용한 한국어 자연어처리: 개체명 인식, 감성분석, 의존 파싱, 의미역 결정," *한국정보과학회 학술발표논문집*, Vol. , pp. 584–586, 2019. [Online]. Available: http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE08763261

[8] 장두성 외, "ALBERT를 이용한 한국어 자연어처리: 감성분석, 개체명 인식, 기계독해," *한국정보과학회 학술발표논문집*, Vol. , pp. 332–334, 2020. [Online]. Available: http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE09874430

[9] M. Kim, J. Byun, C. Lee, and Y. Lee, "Multi-channel cnn for korean sentiment analysis," *Annual Conference on Human and Language Technology*, pp. 79–83, 2018.

[10] C. Park, G. Kim, H. Kim, and C. Lee, "Contextualized embedding-based korean movie review sentiment analysis," *Annual Conference on Human and Language Technology*, pp. 75–78, 2018.

[11] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds., pp. 3111–3119, 2013.

[12] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation." *EMNLP*, Vol. 14, pp. 1532–1543, 2014.

[13] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," *Proc. of NAACL*, 2018.

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Jun. 2019. [Online]. Available: https://www.aclweb.org/anthology/N19-1423

[15] K. Yang, D. Lee, T. Whang, S. Lee, and H. Lim, "Emotionx-ku: Bert-max based contextual emotion

classifier," *CoRR*, Vol. abs/1906.11565, 2019. [Online].
Available: http://arxiv.org/abs/1906.11565