Linguistic Bias Detection and Its Application in Fake News Detection

학번: 22160022
이름: 이재훈
이메일: jm0522@gmail.com

IMLL
Industrial Machine Learning Lab

1. Introduction
2. Methodology
3. Dataset
4. Experiment & Result
5. Conclusion
6. Reference

IMLL
Industrial Machine Learning Lab

1. Introduction

❖ Neutral Point of View(NPOV)

- 위키피디아의 객관적/중립적 정보전달을 위한 글쓰기 정책
- 본 연구에서는 NPOV를 위반한 문장을 탐지하고자 함
- NPOV 분류기를 통해 Fake News 탐지에 활용함
- 예문)

Evolution **is** the source of the vast diversity of extant and extant life on Earth.

Evolution **may be** the source of the vast diversity of extant and extinct life on Earth.

2. Methodology

❖ Data Augmentation

- 단어 사이에 랜덤하게 문장부호를 삽입
- 유의어 대체, 위치변경, 단어 삽입, 단어 삭제

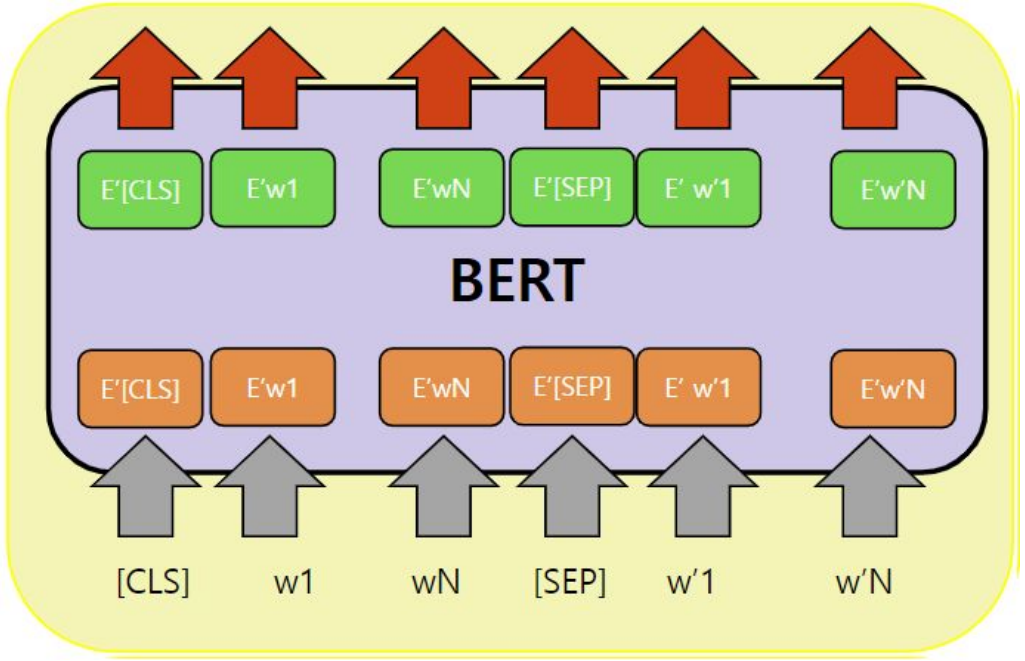
```
✓ eda('no, i do not want it')

['no i do non not want it',
 'no not do i want it',
 'no i do not privation it',
 'no i do not deprivation it',
 'no i do want it',
 'no i do not desire it',
 'no i want it',
 'no i not want it',
 'no i do not want non it',
 'no i do not want it']
```

	sentence	label
0	While this can ; be structured ; like ! a sale...	1
1	While this can be structured like a pure sales...	0
2	He is the father ; of : comedienne Carlen Altm...	1
3	He ; is the father . of comedian Carlen Altman .	0
4	Japan has a : thriving ! fetish scene . with t...	1

2. Methodology

❖ Bidirectional Encoder Representations from Transformers(BERT)

- 사전학습 모델인 **BERT**를 파인튜닝하여 분류모델로 활용



2. Methodology

❖ T-test

- 두 집단의 평균을 비교하는 모수적 통계방식
- 실험결과가 유의미한 차이가 있는지를 보일 때 활용

3. Dataset

❖ WIKIBIAS corpus

- 위키피디아로부터 NPOV위반 문장의 수정기록들을 수집한 데이터셋
- WIKIBIAS corpus 데이터 중에서 주석자들을 통해 필터링된 데이터 8000개와 필터링되지 않은 데이터 10000개 활용

sentence	label
Evolution is the source of the vast diversity of extant and extinct life on Earth.	1
Evolution may be the source of the vast diversity of extant and extinct life on Earth.	0

3. Dataset

❖ Fake News Dataset

News	Size (Number of articles)	*Subjects*	
Real-News	21417	**Type**	**Articles size**
		World-News	*10145*
		Politics-News	*11272*
Fake-News	23481	**Type**	**Articles size**
		Government-News	*1570*
		Middle-east	*778*
		US News	*783*
		left-news	*4459*
		politics	*6841*
		News	*9050*

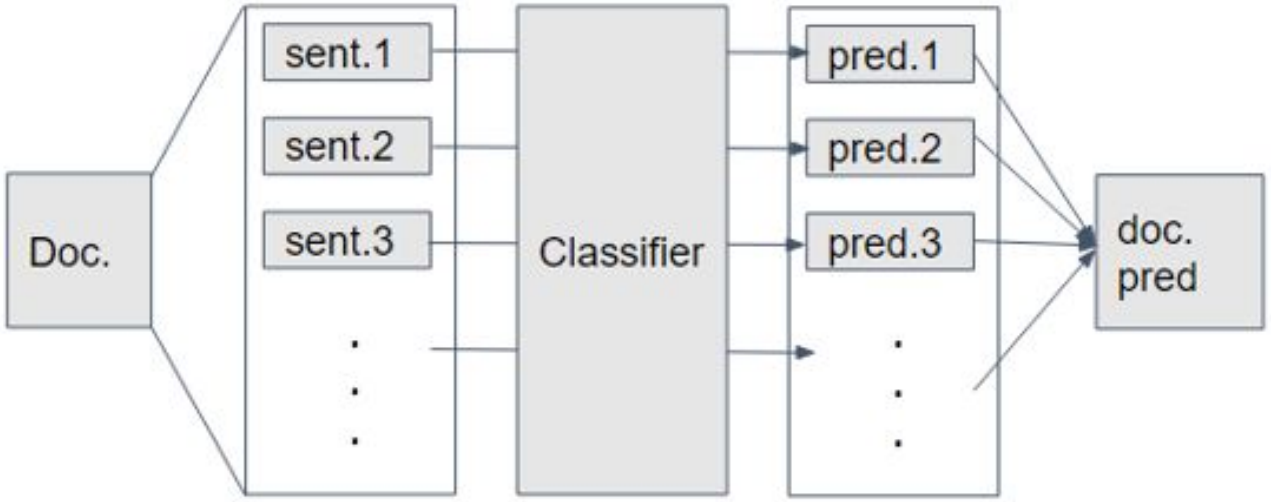
4. Experiment & Result

❖ Bias 여부 분류 모델 성능

model	label	precision	recall	f1	support	acc.
Base model[3]	-	0.70	0.39	0.52	-	0.68
BERT	0	0.70	**0.87**	**0.78**	1252	0.70
	1	0.71	0.46	0.55	852	
BERT_aug	0	**0.76**	0.69	0.72	1252	0.69
	1	0.60	**0.67**	**0.63**	852	
BERT_aug_noise	0	0.72	0.84	0.77	1252	0.71
	1	0.69	0.52	0.59	852	

4. Experiment & Result

❖ 문서의 편향성 계산 Framework



1. 문서내의 문장들을 분류
2. 문서별로 문장들끼리 예측값의 평균을 구하여 문서에 대한 예측값으로 설정

4. Experiment & Result

❖ Result

- Fake news 와 real news의 문서 편향성 통계

	Fake News	Real News
평균	**0.545**	**0.169**
표준편차	0.23	0.14

- Fake news가 평균적인 문서 편향성이 높다는 것을 t검정을 통해 확인 가능

5. Conclusion

❖ 결론

- 데이터 증강기법을 활용하여 기존 베이스라인 모델보다 더 좋은 성능을 기록함
- 이를 Fake news탐지에 활용하여 real news와 fakenews 사이에 유의미한 차이가 있는 것을 확인함

6. Reference

[1] Wikipedia: Neutral point of view. (16 April 2022). Retrieved April 28, 2022, From Wikipedia: https://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view.

[2] Recasens, M., Danescu-Niculescu-Mizil, C., & Jurafsky, D. (2013, August). Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1650-1659).

[3] Zhong, Y., Yang, J., Xu, W., & Yang, D. (2021, November). WIKIBIAS: Detecting Multi-Span Subjective Biases in Language. In *Findings of the Association for Computational Linguistics: EMNLP 2021* (pp. 1799-1814).

[4] Karimi, A., Rossi, L., & Prati, A. (2021, November). AEDA: An Easier Data Augmentation Technique for Text Classification. In *Findings of the Association for Computational Linguistics: EMNLP 2021* (pp. 2748-2754).

[5] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *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).

[6] Ahmed H, Traore I, Saad S. (2017) “Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques. In: Traore I., Woungang I., Awad A. (eds) Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments. ISDDC 2017. Lecture Notes in Computer Science, vol 10618. Springer, Cham (pp. 127- 138).

Thank you

IMLL
Industrial Machine Learning Lab