텍스트 스타일로 MBTI 예측 Predicting MBTI Personalities through Text Styles

유승욱(소프트웨어학부), 김상렬(컴퓨터공학부), 김중훈(응용통계학부), 박경빈(소프트웨어학부)

2021 CUAI 중앙대학교 인공지능 학회 하계 컨퍼런스
Proceeding of 2021 Chung-Ang University Artificial Intelligence's Summer Conference

Abstract

성격 유형 지표로서 사람의 특성을 나타내는 MBTI로
텍스트 분류를 수행했다. 국내 대규모 MBTI 커뮤니티인
MBTI & HEALTH 심리 네이버카페에서 크롤링을 통해
약 44,000개의 MBTI 관련 텍스트를 수집하였다.

많은 종류의 MBTI로 세세한 분류를 진행하는 것에
어려움이 있을 것으로 판단하여 MBTI 열여섯 가지가 아닌,
MBTI를 결정하는 네 가지 지표를 분류하는 라벨링을 진행했다.

머신러닝으로는 Random Forest, XGBoost, Linear/Nonlinear SVM을
딥러닝으로는 LSTM, Bi-LSTM, GPT2 모델을 선택해
각 모델에서 텍스트 분류 모델을 구현하고 그 성능을 확인했다.

이 텍스트 분류 모델을 웹에 바로 띄워주는 프로토타입을 제작해
새롭게 작성된 텍스트로 MBTI 예측 결과를 실시간으로 확인할 수 있다.

Introduction

에너지 방향/인식/판단/생활양식 네 가지 지표를 16개 유형으로 조합!
내 모습을 발견하며, 나와 다를 수 밖에 없는 타인의 모습도 이해!

위와 같이 지표 조합에 따라 사람의 특성을
16가지로 구분하는 MBTI는 현재 젊은 층 사이에서 그야말로 '열풍'이다.
내가 어떤 사람인지 알아가려는, 또 상대방을 이해하고 싶다는
욕구가 반영되며 수많은 누리꾼들의 입소문에 오르내린 것이다.

온라인 매체에서 작성된 텍스트로 MBTI를 예측할 수 있다면
글쓴이 시선에서 작성된 글을 바라볼 기회가 생긴다.
상대방의 관점에서 텍스트 맥락을 이해하며 서로의 글에
공감하는 상황은 곧 원활한 의사소통으로 이어질 것이다.

또한 나의 글로 파악된 MBTI가 내 실제 MBTI와 서로 다르다면
평소 본인의 작문 성향을 파악해볼 수 있다.
본인의 성격이 반영되지 않는 글을 작성한다는 사실과,
실제 글에 반영된 성격이 무엇인지 깨우치게 함으로써
자신의 텍스트 스타일 및 작문 방법에 대한 성찰이 가능해진다.

Dataset

국내 대규모 MBTI 커뮤니티인 MBTI & HEALTH 심리 네이버카페를
선정해 11개의 게시판에서 약 44,000개의 텍스트를 크롤링했으며,
그 과정에서는 Selenium과 BeautifulSoup이 활용되었다.

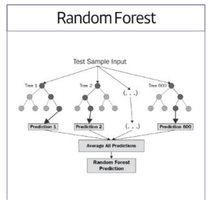
MBTI & HEALTH 심리 카페
CAFE.NAVER.COM/MBTICAFE

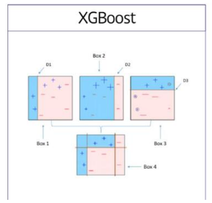
데이터 선정에서는 아래의 고려사항을 반영하였다.
첫째 - 서로의 MBTI를 염두하며 대화가 진행하는지,
둘째 - 얻은 텍스트가 모델링에 유의미하게 적용될지,
셋째 - 충분한 기간 동안 쌓인 데이터를 모을 수 있을지.

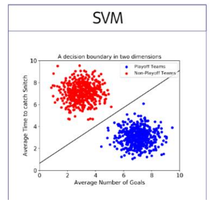
데이터 라벨링에서는 실제 게시판 글을 살펴본 결과
모든 글에 작성자의 MBTI가 그대로 반영되지 않았을 뿐더러
실제 반영되었다 한들 많은 MBTI 종류로 세세한 판단이 불가능했기에
16가지 MBTI가 아닌, MBTI를 구분 지표 4가지로 라벨링을 진행했다.
예시) ENFP의 글을 ENFP라 두지 않고 e_i=0, n_s=0, f_t=0, j_p=1 총 네 가지로 라벨링

또한 정규표현식으로 텍스트 전처리 및 정규화를 통한 텍스트 정제 등의
전처리를 진행하고, '엔프제' 등 MBTI를 지칭하는 표현이 한글로
작성된 경우 'ENFJ' 등 원래의 영어 명칭으로 바꾸는 작업을 거쳤다.

Methods

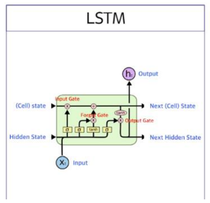


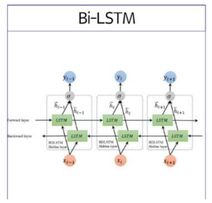


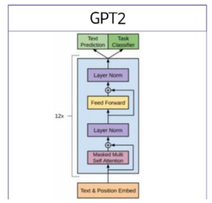


머신러닝 모델로 Random Forest, XGBoost, SVM을 선택하였다.
Random Forest와 SVM에서는 TF-IDF와 Okt로,
XGBoost에서는 Tensorflow Tokenizer로 추가 전처리를 진행했다.

Random Forest는 여러 결정 트리 분류기가 개별 학습을 수행하고
최종적으로 모든 분류기가 Voting을 통해 예측을 결정한다.
XGBoost는 Random Forest와 원리가 동일하나
Boosting을 사용해 오답에 가중치를 부여해가며 예측을 결정한다.
SVM은 데이터 분류를 위해 클래스들 사이의 Margin을 최대화하는
초평면을 찾으며, 선형과 비선형 모델로 종류가 나뉜다.





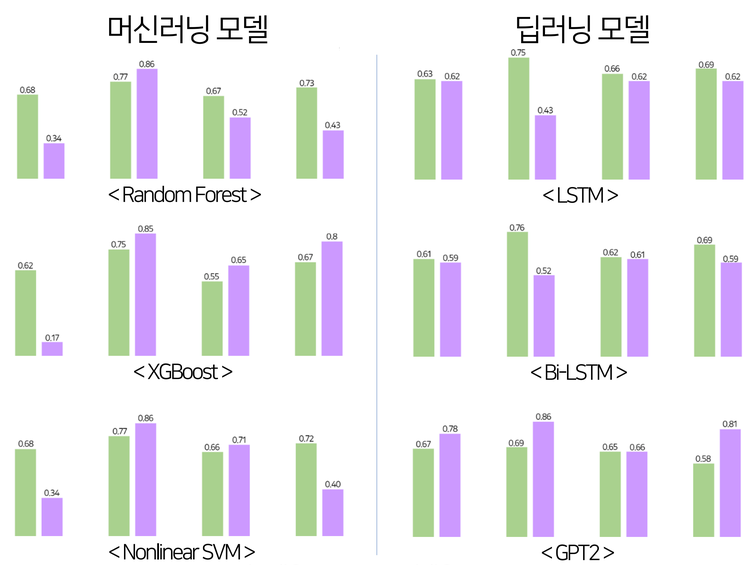


딥러닝 모델로 LSTM, Bi-LSTM, GPT2를 선택하였다.
LSTM 계열에서는 형태소 분석기와 정수 인코딩 및 패딩을 거쳤으며
GPT2에서는 BPE(Byte Pair Encoding)으로 추가 전처리를 진행했다.

LSTM은 기존 simple RNN이 가졌던 장기기억 손실 문제를
극복한 모델로서 그 과정에는 forget gate가 중요하게 작용한다.
Bi-LSTM은 일반 LSTM과는 달리 양방향 모델로서
정방향과 역방향 두가지 순서 모두에서 텍스트 파악이 가능하다.

GPT2는 다양한 도메인의 테스트로 대규모 사전 학습이 진행된
모델로서 기존의 모델들보다 더 많은 영역의 글을 이해할 수 있다.

Results



왼쪽부터 E(외향형) - I(내향형), N(직관형) - S(감각형), F(감정형) - T(사고형), J(판단형) - P(인식형) 판단

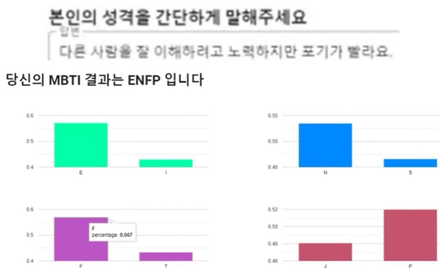
모든 모델에서 네 가지 지표 중 N(직관형) - S(감각형)을 잘 구분했으며,
나머지 지표들은 비슷한 성능으로 분류된 것을 확인할 수 있다.

전반적으로 딥러닝보다 머신러닝 모델들의 성능이 조금 더 높았는데,
이러한 점에 있어 전체 데이터 개수가 영향을 주었으리라 생각했다.
학습한 딥러닝 모델 깊이가 상당한 만큼 더욱 많은 텍스트 레이블
데이터가 있었다면 딥러닝 모델의 성능이 더욱 높았을 것으로 기대된다.

Prototype

학습한 모델을 웹에서
실시간으로 적용하였으며
cuaimbti.site 에서 이용가능하다.

Front-end: vue.js
Back-end: fastapi
Server: aws 및 nginx



Reference

Breiman, L. "Random Forests" Machine Learning 45, 2001
Tianqi Chen and Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System" In Proceedings of the KDD '16, 2016
M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt and B. Scholkopf. "Support vector machines," in IEEE Intelligent Systems and their Applications, vol. 13, no. 4, July-Aug. 1998
Hochreiter, S., & Schmidhuber, J. "Long short-term memory" In Proceedings of the Neural Computation, 1997
Chenbin Li, Guohua Zhan, Zhihua Li. "News text classification based on improved Bi-LSTM-CNN" In Proceedings of 9th International Conference on Information Technology in Medicine and Education (ITME), 2018
Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever. "Language models are unsupervised multitask learners", OpenAIblog, 2019