한동대학교 유정섭 (발표자)
숙명여자대학교 이지수

도입부

발표 시작하겠습니다.

안녕하세요, 11팀 유정섭, 이지수 입니다.

여러분 혹시 MBTI 를 아시나요?

저를 비롯한 많은 분들이 일상 생활에서도 많이 접하실 것이라고 생각합니다.

저는 주로 새로운 사람을 만날 때 첫인상으로 MBTI를 추측하곤 합니다. 하지만 자주 틀려 분위기가 싸해지는 경우를 많이 경험했습니다.

첫인상으로 본 mbti 추측은 정확하지 않았는데 유저가 작성한 텍스트 데이터를 기반으로 MBTI를 예측한다면 더욱 정확한 추측이 가능하지 않을까요?

데이터 소개

저희에게 주어진 데이터는 인터넷 사용자가 작성한 게시글에서 랜덤 추출한 400개의 단어와 사용자의 MBTI입니다.

데이터는 총 74357개였으며, 총 16개의 서로 다른 MBTI들이 있었습니다.

중복 데이터와 결측치를 확인해본 결과 존재하지 않아 따로 처리는 하지 않았습니다.

저희 팀은 텍스트 데이터를 사용해 본 경험이 적어 제공된 캐글에서 래퍼런싱을 하고 접근 방식을 공부하며 이해하는 작업을 했습니다.

먼저, 분석의 큰 틀을 정하며 MBTI의 네 글자를 하나로 묶어서 예측하는 방법과 MBTI 글자 하나하나를 따로 예측하는 모델 4개를 만드는 방법/ 이 두가지를 고려해보았습니다. 하지만, 글자를 따로 학습하는 모델은 전체를 묶어서 학습하는

것보다 현저히 낮은 validation 정확도를 확인할 수 있었기에 전체를 학습하는 것으로 결정하였습니다.

저희가 해결해야할 문제가 다중 분류 문제이기 때문에 LinearSVC / 나이브베이즈 분류기 / 로지스틱회귀 분류기 / 랜덤포레스트 분류기 / 코사인유사도를 이용하는 방법 등 다양한 방법론과 모델을 사용하여 정확도를 확인했습니다.

그 중 비교적 높은 정확도를 보여준 2개의 모델인 LinearSVC와 로지스틱회귀 분류기를 집중적으로 사용하여 진행했습니다.

모델을 선정한 후엔, 세부적인 하이퍼파라미터 조정은 그리드서치를 이용하였고, 그 결과를 바탕으로 전체 training set을 학습시켜 예측값을 얻었습니다.

최종적으로 저희가 사용한 모델은 다음과 같습니다. 저희는 크게 두가지 모델을 사용했는데요, 먼저 자연어처리는 Tf Idf Vectorizer를 사용하였습니다. 저희가 처리방법으로 Tf Idf Vectorizer를 사용한 이유는 CountVectorizer에서는 의미가 없지만 빈도가 높은 단어의 가중치 증가의 **위험**이 있었는데 이를 해결할 수 있는 방법이기 때문입니다.

첫번째 모델은 앞서 말씀드린것과 같이 Tf Idf Vectorizer로 자연어를 처리한 후 하이퍼파라미터 C를 0.3으로 한 LinearSVC모델을 사용하였습니다.

두번째로는, 직접 무의미한 단어를 제거하고, 표제어를 추출한 후 자연어처리를 하는 방법을 사용하였습니다.

두 모델 모두 좋은 성능을 보였기 때문에, mbti 각각의 알파벳을 정할 때, 첫번째 모델을 기반으로 하되, 각 Validation 정확도를 기반으로 0.1이상 차이가 날 경우 두번째 모델의 결과를 사용하였습니다.

(사용 안함)

하지만, 저희가 프로젝트를 진행하면서 MBTI 중에 글에서 가장 잘 드러날 수 있는 성격이 무엇일지에 대해서 생각해보았습니다. 외향적/내향적을 뜻하는 E와 I,,,, 그리고 흔히 계획을 세우는지 안세우는지를 뜻하는 것으로 알려진 J와 P는 글에서 특징이 나타나는 것이 어려울 수 있겠다고 생각한 반면,

직관형인 S와 감각형을 뜻하는 N과, 그리고 감정을 중시하는 F와 결과를 중시하는 T는 글에서 잘 드러날 수 있다고 생각하였고, 실제로 네 글자를 따로 모델링하였을 때 눈으로 확인할 수 있었습니다.

모델 사용 방법 - 활용도

자, 이렇게 생성된 모델은 어디에 사용할 수 있을까요?

저희는 사용자에게 직접 MBTI 를 제공받을 수 없어 유저의 활동으로 예측을 해야하는 상황에 초점을 두었습니다.

첫번째로 고객의 개인화와 만족도 향상을 위한 콜센터에 도입이 될 수 있습니다.

최근 음성 데이터를 텍스트 데이터로 바꾸는 STT (Speech To Text) 기술이 발전되고 있습니다. 축적된 전화 상담 내용을 텍스트로 바꿔 분석해 MBTI 를 예측하고 MBTI별 특징을 살려 고객의 공감대에 맞춰진 상담을 제공할 수 있습니다.

예를 들자면, 친구에게 교통사고가 났다고 전화했을 때 여러분은 어떤 반응을 원하시나요? (발표 시작전 반응 확인하고 좌/우 - 이것처럼 고객마다 원하는 답변이 다릅니다.) 사고형 즉 T 이신 분들은 간단명료하게 앞으로 해야하는 것을 말하는 것을 원하고, 감정형 즉 F 이신 분들은 공감을 원하는 특징이 있습니다. 이에 기반하여 콜센터에서는 고객의 MBTI 를 예측하여 고객 맞춤형 상담을 통해 만족도를 높일 수 있으며, 고객과의 불화를 **방지해** 상담원의 감정 노동의 감소를 기대합니다.

두번째는, 온라인 환경에서 사용자에게 맞춤형 광고와 마케팅에 도움을 줍니다. 각종 SNS와 커뮤니티에 일상을 나누는 글들처럼 다양한 활동들을 하는 것을 볼 수 있습니다. 유저가 작성한 글을 사용하여 유저의 성향을 파악하고 가장 적합한 광고를 보여줄 수 있습니다.

"성격 차이가 광고 효과에 미치는 영향에 관한 탐색적 연구"에 따르면 "개인 소비자의 성격 유형과 광고의 이미지가 일치할수록 광고에 대한 태도는 보다 긍정적이 된다는 시사점을 제시하고"있습니다.

또한, 실제로 관광지 추천을 관광객의 성격유형데이터를 통한 관광동기를 사용하여 관광 유형을 추천해주고 있다는 내용을 많이 찾아볼 수 있었습니다. 관련 연구에 따르면, 관광 동기를 5개의 요인으로 나눴을 때 "모두 관광 동기 부분에서 유의미한 결과를 나타냈으며 이에 따른 관광지 추천이 가능한 것을 시사해주고 있다"고 주장합니다.

이처럼 MBTI 와 마케팅은 밀접한 연관이 있어 광고 개인화에 높은 도움을 줄거라 예상합니다.

결론

MBTI는 개인의 특성과 밀접한 연관이 있습니다. 빅데이터를 통해 개인의 MBTI를 예측할 수 있다면 개인화를 통한 고객의 서비스 만족도 향상과 개인맞춤형 마케팅을 기대할 수 있습니다. 저희는 Linear SVC모델을 사용한 텍스트마이닝을 통하여 사용자의 MBTI를 예측 할 수 있었습니다. 이어서 저희가 만든 예측 모델을 시연하겠습니다.

Run 1: 06:23

Run 2: 06:35

팟팅~~