안녕하세요. 성균관대학교 팀 얼간이 발표 시작하겠습니다. 저희의 주제는 기업 리뷰데이터를 활용한 이직률 예측 모델 제안 및 기업분석입니다.

저희의 연구는 다음과 같이 이루어져 있습니다.

첫번째로 사업 개요입니다. 한국 경제 TV에 따르면, 국내 성장 기업의 3년차 이내 10명 가운데 8명은 퇴사 또는 이직을 고민한 적이 있는 것으로 나타났습니다. 트렌드 코리아에서도 2023년 트렌드 중 오피스 빅뱅을 언급할 정도로 현대 사회는 개인이 우선시 된 시대가 되었음을 알 수 있었습니다. 시대의 흐름이 바뀐 만큼 기업도 이에 대해 대비 해야 합니다. 따라서 저희 팀은 기업별로 이직할 확률과 요인을 머신 러닝을 통해 연구해보자 하였습니다.

해당 주제에 대한 선행 연구로는 한국데이터 과학회에서 진행한 ‘IT 기업 직원의 만족 및 불만족 요인에 따른 이직률 예측’과 대한 산업 공학회에서 진행한 ‘기업 리뷰 텍스트 마이닝을 통한 기업 분석 시스템 개발’을 참고했습니다. 첫번째 논문에서는 이직률을 예측하는 모델을 만들어 기업의 이직률을 예측하였으며, 두번째 논문에서는 이에 대해서 토픽 모델링을 통해 실제로 데이터 분석 후 유효성 검증 등의 데이터 검증 절차를 밟았습니다.

위 논문을 바탕으로 저희 팀의 개발 목표는 다음과 같습니다. 우선 토픽 모델링을 구현하고 포스트 코로나 시대와 MZ 세대들의 취업시장 진입으로 인한 변화를 관측하고자 합니다. 선행 연구에서 얻은 토픽들을 비교하고 실제 기업 데이터를 이용하여 유효성을 검증하고자 합니다. 또한, 궁극적으로 이직률 예측 모델을 개발하여 현재 트렌드를 반영한 기업 별 이직률을 예측하고자 합니다. 저희가 개발한 모델은 이직률 예측을 통해 기업 운영 평가와 방향 수정에 도움이 될 것이며, 각각의 토픽에서 기업의 객관적인 위치를 확인해볼 수 있으리라 생각합니다.

다음으로는 기술 개발 및 결과 입니다. 개발 과정을 순차적으로 보이자면 데이터 수집, 토픽 모델링, 모델 개발 3가지의 부분으로 나누어져 있습니다.

데이터 수집 과정은 다음과 같습니다. 2022년 1월부터 2023년 9월까지의 기업 리뷰 데이터, 재무제표 데이터, 이직률 데이터를 잡플래닛, 원티드 인사이트에서 셀레니움 크롤링을 활용하여 가져왔습니다. 기업 리뷰 데이터의 경우 it, 서비스업, 제조 등 총 10개의 분야에서 100개이상의 리뷰를 가진 회사 중에서 선정해 수집 하였습니다.

가져온 데이터에 대한 전처리 작업을 진행한 후 젠심 라이브러리를 사용하여 LDA 기법을 이용한 토픽 모델링을 진행하였습니다. 해당 그래프는 각 토픽별 응집도와 복잡도를 그래프로 나타낸 것입니다. 토픽 수는 토픽모델링 결과물의 내용이 가장 뚜렷이 구분되는 경우로 정하였으며, 장점 토픽은 총 6개, 단점 토픽은 총 7개로 이와 같이 결정하였습니다. 자세한 내용은 토픽 분석 파트에서 다루도록 하겠습니다.

그 후 데이터의 분표 형태를 살펴 보았습니다. 이직률 데이터 분포와 장점 리뷰 토픽 모델링의 분포가 좌하단에 집중되어 있다는 점을 확인할 수 있었고 단점 리뷰의 분포와 재무제표의 분포 역시 이직률의 데이터 분포와 유사함을 알 수 있었습니다.

이직률 예측으로는 다양한 머신러닝 모델을 사용하여 진행하였으며 머신러닝 모델의 x 값으로는 기업의 이름과 이직률을 제외한 나머지 값을, 그리고 y 값으로는 기업의 이직률을 사용하였으며 평가 지표는 MSE, RMSE, MAE값으로 설정 하였습니다. 그 결과 XGB와 Light GBM 모델이 가장 좋은 성능을 보여줌을 알 수 있었습니다.

학습 시킨 머신 러닝 모델에 대한 피쳐 중요도는 다음과 같이 나왔으며 이를 활용하여 이용가능한 출력물들을 만들었습니다. 먼저 취업 및 이직 준비 인원의 경우 선택한 토픽 별 상위 5개 기업을 검색 해 볼 수 있으며, 기업명을 입력하면 기업 특징에 대한 토픽 모델링값과 이를 다른 기업과 비교한 백분위수로 볼 수 있습니다. 기업 경영자의 경우 기업의 장,단점 리뷰데이터와 재무제표 데이터를 csv 파일로 입력하면 토픽 모델링을 통해 기업의 특징을 확인 할 수 있으며, 학습 시킨 머신러닝 모델로 예측된 이직률도 확인 할 수 있습니다.

다음으로는 데이터 분석입니다.

먼저 장점 토픽에 대해 분석결과를 공유하고자 합니다. 2015년 부터 2018년 데이터를 활용한 선행 논문의 경우, 직원 만족 요인의 토픽으로 유연한 직업 환경, 성과와 인정의 공평함, 잘 설립된 복지 시스템, 만족스러운 급여, 성장 기회, 자유롭게 일하는 환경, 수평 조직 문화,직원 친화적 분위기로 총 8개의 토픽을 선정하였습니다.

선행 논문과 비교하여 새롭게 떠오른 토픽으로 여가 지원을 볼 수 있는데, 이를 보면 직장인들의 여가 생활을 지원해주는 것을 중요한 요인으로 보고 있음을 알 수 있습니다. 잡코리아에서 분석한 Z 세대들의 설문조사에 의하면 Z세대가 가장 궁금해하는 것은 직원의 복지제도로 나왔으며 연구 결과를 뒷받침하고 있습니다.

선행 논문에서도 존재했던 토픽 안의 단어들과 선행 논문 토픽안의 단어들을 비교 분석하였을 때 자율적 근무 환경 토픽의 단어들이 크게 바뀌었음을 알 수 있었습니다. 논문의 자율적 근무환경은 연차, 휴가, 눈치와 같이 휴가를 눈치 보지 않고 사용할 수 있는 환경을 자율적 근무 환경이라고 보았는데, 2023년의 자율적 근무 환경의 단어들을 보면 재택 근무, 재택과 같은 단어들로 재택근무를 허용해주는 회사들을 자율적 근무 환경이 보장된 회사로 봄을 알 수 있었습니다.

이번에는 단점 토픽에 관해 말씀드리겠습니다. 선행 논문에서는 불확실한 성장 전망, 열악한 업무 프로세스와 조직 구조, 사내 정치와 불공정한 보상, 낮은 연봉과 미흡한 복리후생, 꼰대 문화, 불합리한 업무 배분, 보수적인 조직문화로 구분되었습니다. 저희 토픽과 비교하여, 먼저 정규직 전환으로의 어려움은 새롭게 등장한 토픽입니다. 비정규직 인원이 연속 증가하며 작년에 역대 최대치를 기록하였으며, 정규직과 비정규직 임금 격차가 지속적으로 증가하였습니다. 이로 인해 많은 사람들이 정규직 전환에 많은 관심을 가진다는 것을 알 수 있습니다.

고객 응대와 성과에 대한 압박 또한 새롭게 등장한 토픽입니다. 이는 한국노동안전보건연구소에서 발표한 노동강도를 강화시키는 요인들입니다. ‘영업, 업무 성과에 대한 압박’, 민원 및 고객대면업무에 대한 스트레스’가 가장 크게 나타난 것을 보아 해당 토픽이 직업을 고르는 과정에서 매우 중요시되는 부분임을 알 수 있습니다. 따라서 해당 토픽에 대한 증가된 관심이 토픽모델링의 결과에 반영되었다 생각합니다.

토픽 모델링에 대한 유효성을 알아보기 위해 실제 기업 중 아모레 퍼시픽의 기업 리뷰를 토픽 모델링 하여 나온 결과 값을 확인하였습니다. 아모레퍼시픽의 경우 복지 시스템, 자율근무 환경, 정규직 전환, 고객응대와 같은 토픽에서 높은 값을 보이는 것을 확인할 수 있었습니다. 실제로 자율 근무 환경과 관련해서 MBC에서 진행한 아무튼 출근! 이라는 프로그램에서 아모레 퍼시픽 직원은 출근 후 업무 시작 전 자유시간을 보장 받고 탄력 근무제로 인해 조기 퇴근 하였으며, 이에 대해 만족하는 모습을 보였습니다.

머신러닝의 피쳐 중요도의 결과는 요약하자면 다음과 같습니다. Light GBM과 XGB 둘 다 평균 연봉과 수당 + 회사시설 및 제도의 열악함에 높은 중요도를 부과했으나, 그 이후 피쳐들에 대해서는 차이를 보였습니다. Light GBM의 경우 자율적 근무 환경, 급여, 정규직 전환 등에 높은 중요도를 부과했지만, XGB는 보수적 문화, 복지시스템, 직원 친화적 분위기에 높은 중요도를 부과했습니다.

이제 기술 활용 방안으로 넘어가보겠습니다.

해당 기술은 일반 취업 및 이직 준비 인원과 기업 경영자의 관점에서 활용 가능합니다. 전자의 경우, 선택한 회사의 토픽 수치를 백분위로 확인하여 각 토픽별로 타 기업 대비 만족도 및 불만이 어느정도 되는지 확인 할 수 있습니다. 또한, 자신이 중요하다고 생각하는 토픽을 입력해서 해당 토픽에서 상위권 회사를 찾을 수 있으며, 자신의 취업 플랜에 이용해볼 수 있습니다. 후자의 경우, 인력 수급에 대한 계획과 운영에 활용해 볼 수 있으며, 경쟁회사와의 비교를 통해 회사 성장에 도움이 되도록 전략 수립에 도움을 줄 수 있습니다. 이는 장기적 관점에서 직원의 이탈을 줄일 수 있으며, 인사 및 직원 교육에 필요한 비용 절감을 기대해 볼 수 있습니다.

해당 기술을 활용 가능한 예시로는 아마존의 펄스 서베이가 있습니다. 아마존은 매일 아침 업무를 시작하기 전에 조직에 관한 질문에 답변을 하도록 요청했습니다. 설문 답변을 기반으로 아마존은 2015년 최악의 조직문화를 가지고 있던 회사에서 2018년 미국에서 가장 바람직한 조직 문화를 갖고 있는 회사로 평가가 바뀌게 됩니다. 설문 그 자체도 즉각적으로 활용 가능하지만, 이를 모델에 활용해본다면 숨어있는 기업의 특성까지 찾아볼 수 있으리라 생각됩니다. 또한, 모델 가중치에 설문 내용을 재학습 시켜본다면 지금 현재만의 트렌드를 예측하는 것이 아닌 장기적으로도 활용이 가능할 것입니다.

마지막으로 한계점과 후행 연구 제안으로 넘어가고자 합니다.

저희가 연구에서 찾은 한계점은 2개입니다. 먼저, 검색 조건에 해당하는 기업의 개수가 부족하였기에 머신러닝 모델이 학습을 충분히 하지 못했던 것 같습니다. 그렇기에 국내 데이터에 한정되지 말고 해외 기업 데이터를 활용 하거나 해외로 시선을 바꾸는 방향도 제안하고자 합니다. 다른 하나의 한계점으로는 사용한 웹사이트 별 기업의 정보에 있어서 불일치가 발생했으며 이로 인해 추가적인 데이터 손실이 발생했습니다.

다음으로는 후행 연구 제안입니다. 먼저, 토픽 모델링을 활용해, 이직 희망자의 직장의 단점 특성을 분석해서 비슷한 직종에서 단점을 보완하는 회사들을 추천하는 추천 시스템을 연구를 제안하고자 합니다. 해당 연구는 이직 희망자의 니즈를 충족 시키면서 사회적 단절을 막는 역할을 해줄 수 있으리라 생각됩니다. 그 다음, 현 연구에서 찾은 발견점들을 반영하기 위한 기업의 전략을 파악한 후 5에서 10년 후 시행했던 기업들의 성장 여부 파악하는 연구도 제안하고자 합니다. 마지막으로 토픽 모델링 시 단점과 장점 토픽에서 자주 나왔던 “소통, 보수, 꼰대, 복지”와 같은 주요 단어들에 대한 근본적인 해결책을 제안하는 후속연구를 경영학과 통계학에서 진행해보면 좋으리라 생각됩니다.(이 부분은 한 가지 중요한 거만 찝어주는 게 어떤지 시간이 부족하다면)

긴 발표 들어주셔서 감사하고 마치도록 하겠습니다. 질문이 있으시다면 자유롭게 해주시면 감사하겠습니다.