SNS 한국어 텍스트 기반의 학습데이터 구축

- 학습데이터 품질향상을 위한 라벨링 애플리케이션 개발

팀 : 미니언즈

강승범, 김인용, 노규명, 조소정, 조하담, 하은혜, 황지상

목차

1. 프로젝트 개요

4 차 산업혁명시대를 앞둔 지금, 현 시대를 대표할 수 있는 키워드는 '데이터'이다. 4 차 산업혁명에서의 데이터는 기존의 숫자나 범주형 자료에 의존하는 정형 데이터 뿐만 아니라 그동안 데이터로서의 가치가 없었던 이미지, 텍스트 등 비정형 데이터를 포함한다. 다양한 형태의 데이터와 분석법의 신속한 개발로 구축된 빅데이터는 이미 우리 생활 속에 자리잡아있다. 개인의 시청기록을 분석해 취향에 맞는 영화를 추천해주는 알고리즘, 인터넷 쿠키를 이용한 장바구니 분석 등 다양하게 사용된다. 데이터의 활용성이 다양화됨에 따라, 챗봇을 이용한 고객관리 서비스, 욕설을 필터링하는 악플감지시스템 등 텍스트를 데이터로 이용한 텍스트 마이닝 사례 또한 증가하고 있다.

텍스트 분석의 성능은 올바른 모델링과 방대한 양의 학습데이터 구축에 의해 좌우된다. 특히, SNS 텍스트의 경우 유행을 반영하는 언어가 주로 사용되기에 변화 주기가 빠른 특징을 보인다. 빠른 변화주기에 대응하려면 학습데이터의 지속적인 업데이트가 필요한데 이를 가능하게 할 효과적인 학습데이터 구축방법을 만들고자 한다. 학습데이터 구축을 위해서는 텍스트에 대한 라벨링이 필요하다. 일반적으로 라벨링은 사람이 일일히 해야하는 작업이기에 지루한 절차로 여겨진다. 그래서 라벨링에 오락적인 요소를 가미하여 게임화하는 gamification[1]의 방법으로 기존의 따분한 라벨링 과정을 일이 아닌 놀이로 바꾸어 관심을 유도하면 학습데이터의 양이 빠르게 증가하고 이는 분석 성능 향상으로 이어질 것이라 생각했다. 이에 대한 방안으로 라벨링 애플리케이션을 제안한다.

인스타그램은 현재 우리나라에서 가장 보편적으로 이용되는 SNS 중 하나로 20%의 점유율을 차지하고 있다. 다수의 SNS 플랫폼 중 유일하게 매년 성장세를 보이는 플랫폼이기 때문에 학습데이터의 미래 활용성을 고려해 여러 SNS 중 인스타그램을 텍스트 데이터로 선택했다.

2. 선행연구

인스타그램의 텍스트를 데이터로 이용하고자 한 연구는 활발히 진행되고 있다. <*인스타그램 기반 이미지와 텍스트를 활용한 사용자 감정정보 측정(2014)*>[2] 은 인스타그램의 이미지와 텍스트(영어)를 이용해 이용자의 감정상태를 분석했다. 이미지의 대표 색상을 추출해 색상에 맞는 감정 형용사를 댓글과 비교한 후 포스트를 대표하는 감정 형용사를 추출하는 연구이다. 그 결과 이미지만 분석했을때보다 이미지와 텍스트를 함께 분석하였을 때 예측 정확도가 더 올라가는 것이 확인됐다.

SNS 텍스트를 통해 프로필을 예측하고자 한 연구도 존재한다. <*Twitter user profiling based on text and community mining for market analysis(2013)*>[3] 는 트위터 텍스트를 데이터화하여 이용자의 프로필을 예측했다. 연구에선 연령대, 성별, 거주지, 직업, 기혼여부를 텍스트 데이터를 통해 유의하게 구별하였다. 이 결과를 통해 텍스트를 통해 이용자의 프로필 예측이 가능함을 알 수 있다.

앞서 소개한 연구들을 통해 SNS의 텍스트를 이용한 데이터의 유효성이 증명되었다. 다만, 외국어를 기반으로 했기 때문에 한국어 텍스트에서의 성능을 보장할 수 없다. 그리고, 후자의 연구는 인스타그램이 아닌 트위터를 데이터로 사용했다는 점에서 차이가 존재한다. 트위터는 텍스트 기반의 SNS 플랫폼이지만 한국에선 정치적 혹은 개인의 취미 등 목적성이 뚜렷한 경향이 있어 이용자가 특정계층에 편향될 가능성이 존재하므로 트위터 텍스트는 우리의 프로젝트에서 데이터로 사용하지 않았다.

---

[1] 게임이 아닌 분야의 문제를 게임의 요소를 도입해 해결하는 방식

[2] 남민지, 김정인, 신주현

[3] Kazushi Ikeda, Gen Hattori, Chihiro Ono, Hideki Asoh, Teruo Higashino

3. 학습데이터 구축

3.1. 라벨링을 위한 데이터 수집

학습데이터 구축을 위해선 가장 먼저 라벨링을 할 데이터를 수집해야 한다. 임의로 데이터를 수집하는 경우 이용자에 대한 편향이 생길수 있기 때문에 사전에 카테고리를 정의하고 분류별로 약 7천명의 이용자를 추출했고 그 이용자들의 포스트를 수집했다. 카테고리를 성별, 연령대, 지역, 직업, 관심사로 나누었고 카테고리를 대표할 수 있는 키워드를 해시태그 검색으로 이용자를 추출했다. 다만 지역 카테고리의 경우, 지역 거주민이 아닌 방문자도 포함이 됐기 때문에 해시태그기반 검색이 아닌 지역이 확실한 몇 유저를 선택해 그의 커뮤니티를 분석하는 방법으로 데이터를 수집했다. 또한 광고를 제외하기 위해 이용자의 팔로잉, 팔로우, 포스트 개수에 임계치를 설정해 임계치를 넘는 이용자는 수집에서 제외했다.

카테고리	카테고리 분류	해시태그 검색 키워드
성별	남성	남중, 남고, 군대, 입대, 신병휴가, 전역 등
	여성	여중, 여고, 여대, 이화여대, 숙명여대, 성신여대 등
연령대	미성년자	초1, 초4, 초딩스타그램 등
	20대	20대, 개강, 대학생, 대학생공스타그램, 학사모 등
	30대	30대, 계란한판, 서른, 스물열살 등
	40대 이상	40대, 마흔, 마흔틴, 50대, 50대몸짱, 50대아줌마, 꽃중년 등
직업	경찰	경찰, 경찰스타그램, 중앙경찰학교 등
	교사	교사, 교사스타그램, 쌤스타그램 등
	군인	군스타그램, 군인, 군인스타그램 등
	소방관	소방공무원, 소방관, 소방스타그램 등
	의사	의사, 의사스타그램 등
	간호사	간호사, 간호사그램, 나이팅게일선서식, 널스타그램, 응급실간호사 등
	중고생	중1, 중2, 중3, 고1, 고2, 고3 등
관심사	게임	롤, 메이플스토리, 모바일게임, 게임스타그램, 게임추천 등
	맛집	맛스타그램, 맛있다, 먹방, 먹스타그램, 맛집그램 등
	미디어감상	유튜브, 유튜버, 구독, 넷플릭스, 미드, 영화스타그램 등
	반려동물	반려견, 반려묘, 강아지, 댕댕이, 펫스타그램 등
	스포츠	운동, 등산, 라이딩, 자전거, 축구, 야구, 헬스, 서핑 등
	여행	여행, 여행스타그램, 여행에미치다 유디니 등
	카페	카페스타그램, 카페투어, 감성카페, 카페추천, 커피맛집 등
	패션	오늘뭐입지, ootd, 오오디티, 남자코디, 여자코디 등

Table 1 : 분류별 해시태그 검색 키워드

3.2. 라벨링 애플리케이션

수집한 데이터의 효과적인 라벨링을 위해 라벨링 전용 애플리케이션을 개발했다. 라벨링 애플리케이션의 이름은 'Whale'과 'Labeling'을 합친 'Whaleling' 으로 했다. 방대한 양의 국어 말뭉치를 엄청난 크기의 바다에 비유하여 하나의 라벨링 과정을 바다에서 해산물을 얻는 낚시에 비유했다. 바다에서 얻을 수 있는 최고의 해산물인 고래를 애플리케이션의 주요 테마로 잡아 최고의 학습데이터를 구축하고자 하는 의지를 바다에서 고래를 찾고자 하는 어부에 대입하였다.

Figure 1 : Whaleling 초기화면

애플리케이션 내에 라벨링의 기능만 존재한다면, 기존의 지루한 라벨링 작업과 차이가 없으므로 추가적인 기능을 넣었다. 먼저, 라벨링 작업을 한 만큼 포인트를 부여했다. 그리고 누적된 포인트를 이용해서 애플리케이션 이용자가 자신의 인스타그램 아이디와 지인의 인스타그램 아이디의 프로필을 검색해보는 기능을 추가했다. 검색한 인스타그램 아이디의 프로필을 예측해 결과창에 보여주고 애플리케이션 이용자가 오류를 수정할 수 있게 했다. 수정된 오류는 다시 학습데이터로 들어가 학습데이터의 지속적인 업데이트가 가능토록 했다.

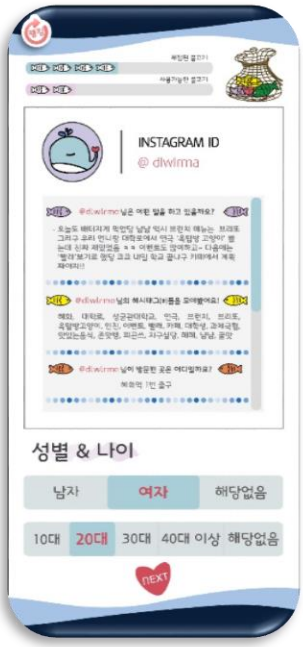

Figure 2 : 라벨링 화면

Figure 3 : 아이디검색 화면

Figure 4 : 분석결과 화면

애플리케이션 상에서의 프로필 분석은 SVM 모델을 이용해 분류하는 텍스트 기반 분석으로 결과를 예측했다. 다만, 포스트 개수가 적어 충분한 양의 텍스트가 없는 경우, 이용자의 팔로잉과 팔로우끼리 같은 카테고리를 공유하고 있다고 전제해 팔로잉과 팔로우의 텍스트를 분석하는 커뮤니티 분석을 병행했다.

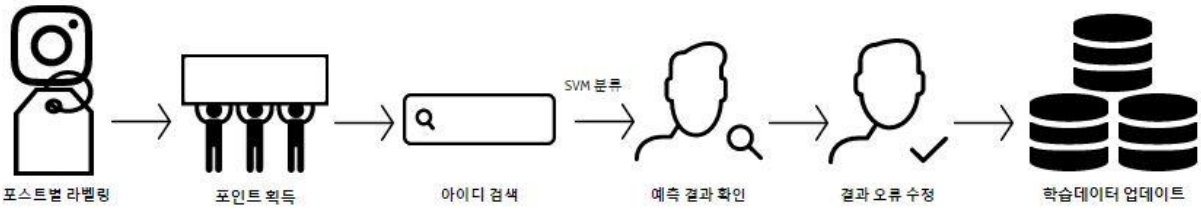

Figure 5 : 애플리케이션 구현 방식

4. 결론

4.1. 활용방안

이번 프로젝트는 완벽한 모델이 아닌 MVP모델[4]을 제안하였다. 최소의 기능으로 필요한 부분만 구성하여 기획했기 때문에 당장의 성과는 기대하기 힘들다. 다만, 학습데이터의 정확성을 높이기 위한 방안으로 지속적으로 업데이트 가능한 모델을 제시하였다. 이를 통해 시간이 지날수록 학습데이터의 양은 증가하게 되고 기존의 국립국어원에서 제공하는 모두의 말뭉치와 결합하여 더 방대한 양을 데이터구축이 되면 더욱 활용가치가 상승할 것이다.

애플리케이션 상에서는 이용자의 프로필 분석에 학습데이터를 이용했지만, 학습데이터를 이용하는 방법은 다양하다. 애플리케이션과 같이 프로필을 예측해 타겟 마케팅에도 사용가능하고, 구축된 텍스트들의 감성분석을 통한 새로운 인사이트 도출도 가능하다.

4.2. 프로젝트의 성과와 한계

우리는 빠르게 변화하는 SNS상의 언어를 분석하기 위해 지속 가능한 학습데이터 구축 모델을 만들었다. 또한 지루한 라벨링 과정에 오락적인 요소를 가미하여 재미있는 라벨링이 되고자 애플리케이션을 개발했다. 이용자가 직접 라벨링 한 만큼 인스타그램 아이디를 분석할 수 있도록 적절한 보상을 해 지속적으로 라벨링을 하고자 유도하였고 순위를 보여줌으로써 경쟁심을 유발하였다. 경쟁적인 라벨링 생태계를 애플리케이션 내에서 구축했기 때문에 학습데이터의 구축은 기존의 라벨링보다 효과적으로, 신속하게 증가할 것이며, 데이터 구축에 속도를 부여할 수 있다. 추후, 국립국어연구원과 말뭉치 공개에 대해 협의 후 공개에 문제가 없을시, 보급해 타 연구의 결과에 유의미한 결과를 미칠 것으로 예상된다.

다만, 포스트와 이용자의 편향을 제거하기 위해 분류별로 나누어 해시태그 기반의 검색을 진행했으나 해시태그 기반의 검색으로 편향성을 완벽히 제거할 순 없기 때문에 카테고리별로 고르게 데이터를 수집하지 못했다. 또한 이용자 수와 누적 라벨링 수가 적은 초기에는 활용가치가 떨어지는 한계가 존재한다.

---

[4] 최소기능제품. 고객의 피드백을 받아 최소한의 기능을 구현한 제품으로 유효성을 검증하고 피드백을 모으기 위한 목적