Summary Report





**과 목 명: AI와 빅데이터를 활용한 경제정보처리**

**과목코드: ECO5320**

**담당교수: 양현주**

**이 름: 황상민**

**학 번: 120220243**

**전 공: 경제학과 AI금융전공**

**제 출 일: 2022년 6월 22일**

Content

1. **Motivation**
2. **Clear Research Question**
3. **Machine Learning Approach**
4. **Data Collection and Processing**
5. **Results**
6. **Value-added to field**
7. **Concluding Remarks: Policy Relevance and Limitation**

**Research Proposal for ECO5320 AI and Economic Data**

Name: Sang Min Hwang, Student ID: 120220243

1. **Motivation**

**요즘 신제품을 출시하면 e커머스에 제품을 등록하는 것은 불가피하다. e-커머스 플랫폼에서 사용자는 필터기능보다는 검색기능을 주로 활용한다. 검색은 키워드를 통해 이뤄지는데, 수많은 제품은 여러 페이지에 걸쳐 노출된다. 그중에서 상위 페이지에 노출될수록 상품은 소비자에게 많이 노출되어 상대적으로 하위 페이지에 노출된 상품보다 매출이 높을 수밖에 없다. 수많은 제품을 e커머스 플랫폼에서 상품을 배치하는 것이 상품의 이미지와 어떤 상관관계가 있는지 알아보고자 첫 번째 질문을 선정했다. 두 번째 질문은 매일 생성되는 비정형 데이터인 뉴스에 관한 것이다. 소비자물가지수(CPI)나 재무제표나 이런 지표들은 신뢰성이 높지만, 정보가 제공되는 간격이 길다는 단점이 있다. 그래서 매일 제공되는 뉴스로부터 의사결정의 근거가 되는 정보를 추출한다면 현재와 수집한 데이터 사이의 괴리가 크게 좁혀질 것이라는 기대에 근거하여 두 번째 질문을 선정했다.**

1. **Clear Research Question**

**이 보고서에서는 두 가지의 질문에 대해 다룬다. 첫 번째 질문은 국내 전자 상거래 웹사이트 쿠팡에서 ‘세제’를 키워드로 검색했을 때, 각 상품의 이미지를 기준으로 상품이 노출되는 페이지 번호를 예측하는 것이다. 한 연구에 따르면 제품포장에서 시각이미지는 구매동기를 유발하고 촉진하는 핵심적인 요소이자 상품의 경쟁력 강화를 위한 마케팅 수단이라고 한다.[1] 그래서 이 질문은 상품의 이미지와 노출되는 페이지 번호 사이에 상관관계가 있는지 확인함으로써 쿠팡에 세제를 판매하고자 하는 신규 판매자에게 상위 페이지에 노출되기 위한 상품 이미지 선정 전략을 제안한다. 두 번째 질문은 유튜브 뉴스 데이터를 이용하여 뉴스 카테고리를 분류하는 것이다. 뉴스 음성데이터 및 썸네일을 활용하여 뉴스 카테고리를 분석하는 간단한 질문으로서 개방된 플랫폼에서 음성과 이미지를 추출하여 분석하는 방법을 제안한다.**

1. **Machine Learning Approach**

**첫 번째 문제인 세제 이미지를 기준으로 페이지 번호 예측에서는 이미지 데이터를 활용하기 때문에 pytorch 기반으로 만들어진 fast.ai라는 딥러닝 라이브러리를 이용하여 간편하게 CNN Regression 모델을 적용했다. 그리고 두 번째 문제인 유튜브 뉴스 카테고리 분류는 두 가지 하위문제로 나뉜다. 뉴스 음성을 음성인식 모델을 이용해 텍스트로 변환하여 RNN 기반의 자연어 분류모델로 분류하는 것과 뉴스 썸네일을 이용하여 fast.ai의 CNN Classification 모델로 분류하는 것이다.**

1. **Data Collection and Data Processing**

**첫 번째 문제는 쿠팡 웹페이지의 URL 구조를 이용하여 주어진 키워드 내에서 모든 페이지에 있는 상품의 정보를 찾고, 각 페이지의 HTML 구조 속에서 필요한 정보를 추출하는 웹 스크래핑을 사용하여 데이터를 추출했다. 여기서 중요한 데이터는 CNN Regression을 수행하는 데 필요한 각 상품의 이미지 URL과 페이지 번호였다. 두 번째 문제인 유튜브 뉴스 카테고리 분류를 수행하기 위해 유튜브에서 스포츠 뉴스, 경제 뉴스, 정치 뉴스를 4~20분 길이의 조건을 적용하여 상단에 노출된 순서대로 약 20개씩 수집하여 카테고리별로 재생목록에 수집하였다. 두 번째 문제는 두 가지 하위 문제로 나뉘어서 데이터 수집도 두 가지로 나뉜다. 첫 번째 하위 문제 데이터 수집에서는 음성을 텍스트로 변환하기 위해 pytube라는 라이브러리를 이용하여 유튜브 뉴스 동영상을 다운로드 받는다. 그리고 그 동영상을 MP3형태의 음성데이터로 변환하고, Naver에서 제공하는 음성인식 모델에 input할 수 있도록 1분 단위로 잘랐다. 두 번째 하위 문제에서는 pytube를 이용해 뉴스의 썸네일 URL을 쉽게 수집할 수 있어서 썸네일과 카테고리를 결합하여 fast.ai의 CNN Classification을 적용하였다.**

1. **Results**

**첫 번째 문제는 쿠팡 세제 키워드로 검색했을 때, 이미지를 기반으로 페이지 번호를 예측하는 것이기 때문에 fast.ai CNN Regression을 적용하였다. 성능 기준은 RMSE로 하였고, monitor= ‘valid\_loss’, patience=10을 parameter로 하여 early stopping을 적용한 결과, 약 7.73의 RMSE를 보여주는 모델로 학습시킬 수 있었다. 쿠팡의 각 키워드 검색결과는 최대 27page까지 있는데, 데이터수집 과정에서 오류를 막기 위해 26페이지까지 수집하였다. 그래서 예측 대상인 페이지의 범위가 1~26인데 RMSE가 7.73인 것은 예측값에 대한 1시그마(약 신뢰구간 68%) 구간은 예측값-7.73 ~ 예측값+7.73이다. 이는 상품의 이미지와 페이지 번호 사이에 약간의 상관관계는 존재한다고는 볼 수 있지만, 그리 유의미하지는 않음을 보여준다. 두 번째 문제 중 첫 번째 하위문제인 유튜브 뉴스 텍스트를 이용한 카테고리 분류는 test dataset에 대해 accuracy가 0.3이 나왔다. 결과가 좋지 않은 이유는 적은 데이터 때문에 발생한 것으로 추정된다. 왜냐하면 <그림 1>에 따르면 epoch가 진행되며 발생하는 loss가 training dataset에서는 줄어드는 반면, validation dataset에 대해서는 줄어들지 않아 오버피팅이 발생했기 때문이다. 두 번째 하위문제는 test dataset에 대해 accuracy가 0.7이 나왔다. 이는 유튜브 썸네일로 뉴스 카테고리를 잘 예측할 수 있음을 의미한다. 즉, 언론사 유튜브 담당자들이 뉴스의 썸네일을 어느 정도 뉴스 내용을 예측할 수 있도록 신경 써서 올린다고도 볼 수 있다.**

1. **Value-added to field**

**첫 번째 질문에 대한 대답을 찾아가며 국내에서 규모가 큰 전자 상거래 플랫폼에서 세제 품목을 중심으로 소비행태를 파악했을 때, 상품의 이미지는 소비와 전혀 관련이 없는 것은 아니지만 약간의 상관관계가 있음을 추측할 수 있었다. 이는 쿠팡뿐만 아니라 다른 전자상거래 플랫폼에서 상품 이미지와 페이지 번호의 상관관계를 파악하는 것이 유의미할 수 있음을 의미한다. 이를 통해 상단 페이지 번호에 노출되는 상품 이미지의 특징을 파악하여 이를 모방하는 전략을 세우는 것과 같은 의사결정을 내릴 수 있다. 그리고 두 번째 질문에 답하는 과정에서 누구나 수집 가능한 방법으로 뉴스 음성을 텍스트로 추출할 수 있음을 보여주었고, 유튜브 뉴스 썸네일 이미지에 담긴 정보가 카테고리를 분류할 수 있는 수준의 정보를 담고 있음을 확인했다. 이는 단순한 뉴스 카테고리뿐만 아니라 다뤄지는 객체(기업, 국가 등)에 대한 긍부정을 판단할 수 있는 것과 같이 거시경제 지표 예측이나 관련 투자 상품에 대한 event에 관한 정보도 뉴스 썸네일 이미지가 갖고 있을 수 있음을 시사한다.**

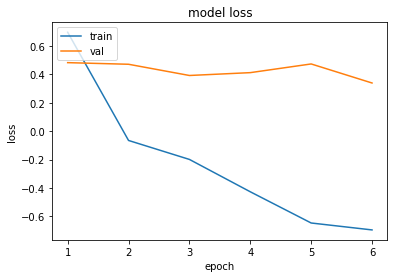
1. **Concluding remarks: Policy Relevance and/or Limitations**

**일련의 Assignment를 수행하며 비정형 데이터를 수집 및 가공하여 기계학습 모델을 통해 예측을 수행했다. 이 예측 결과 자체가 중요하기보다는, 이 결과를 이용해서 경제금〮융에 이용하는 것이 훨씬 중요하다. 첫 번째 질문에서는 단순히 쿠팡에서 이미지를 기준으로 상품이 어떤 페이지를 부여받는지 상관관계를 확인해봤다. 여기에 보태서 이미지와 페이지 번호를 주기적으로 수집함으로써 시계열 데이터를 만들어 코스피와 같은 경제지표와 결합한다면 경기변동에 따른 이미지 특징에 따른 특정 품목의 소비행태를 분석해볼 수도 있다. 두 번째 질문에서는 뉴스라는 매일 제공되는 정보를 유튜브에서 수집하여 카테고리 분류를 수행했다. ‘(기업명 or 국가명) + 뉴스’와 같은 키워드로 특정 기업이나 국가에 대한 뉴스 데이터를 조직명을 레이블로 하여 수집한 다음, 조직의 전망이나 성과를 긍부정으로 예측하는 연구로 확장할 수 있는 주제다.**

**Reference   
(use Harvard style reference,** https://www.citethisforme.com/harvard-referencing**)**

[1] 김경선 and 강혜숙. 2010, "제품포장의 시각이미지에서 인지반응과 감성반응이 제품태도와 구매의도에 미치는 영향", 한국디자인포럼, no.28, pp.197-210.

**Figure and Table Appendix**



<그림 1. [질문2] NLP Classification Model Loss>