한국컴퓨터정보학회 하계학술대회 논문집 제31권 제2호 (2023. 7)

순환신경망(RNN)을 통한 자연어 감성 분석

허태성O, 전세현*
O인하공업전문대학 컴퓨터정보과,
*인하공업전문대학 컴퓨터정보과
e-mail: tshur@inhatc.ac.krO, 201844063@inhatc.ac.kr*

Natural language sensitivity analysis using RNN

Hur Tai-sungO, Jeon Se Hyun*
ODept. of Computer Science Engineering Inha Technical College,
*Dept. of Computer Science Engineering Inha Technical College

● 요 약 ●

본 논문에서는 딥러닝 기법 중 하나인 순환신경망(RNN)을 활용하여 자연어를 처리할 수 있는 모델 개발에 대하여 연구를 진행하였다. 다양한 주제에 대한 사용자들의 의견을 확보할 수 있는 유튜브 플랫폼을 활용하여 데이터를 확보하였으며, 감성 분류를 진행하는 만큼 학습 데이터셋으로는 네이버 영화 리뷰 데이터셋을 활용하였다. 사용자는 직접 데이터 파일을 삽입하거나 혹은 유튜브 댓글과 같이 데이터를 외부에서 확보하여 감성을 분석할 수 있으며, 자연어 속 등장하는 단어의 빈도수를 종합하여 해당 데이터들 속 키워드는 무엇인지를 분석할 수 있도록 하였다. 나아가 종합 데이터 분석 관리 플랫폼을 제작하기 위하여 해당 데이터를 데이터베이스에 저장하고GUI 프로그램을 통하여 접근 및 관리가 가능하도록 하였다.

키워드: 자연어처리(Natural Language Processing), 순환신경망(RNN), 감성분석(Sensitivity analysis)

I. Introduction

인류는 생존을 위하여 조직 생활을 시작하였고, 그 결과 소통의 중요성이 강조되었다. 상대방의 의견을 빠르고 정확하게 파악하는 것은 생존에 있어 중요한 것이며, 특히 최근과 같이 SNS를 통한 쉽고 빠른 의사소통이 가능한 시대에 반드시 필요한 능력이라 할 수 있다. 이러한 수요에 맞춰 글을 인식하고 분석하는 자연어처리 기술은 앞으로도 많이 각광받게 될 것으로 보인다. 과거엔 언어의 모호성, 데이터의 부재 등의 문제점들로 인하여 관련한 시스템의 개발이 어려웠지만, 최근 AI의 등장 및 발전[1]으로 자연어 처리 능력이 개선됨으로써 수월한 개발이 가능해졌다.

II. Preliminaries

1. Related works

1.1 자연어

자연어는 인간과 인간이 상호 소통 시에 사용하는 언어로서, 일상어라고도 불린다. 자연어 처리는 언어의 모호성, 문장 내 정보 부족, 같은 의미의 다른 표현 등 극복해야할 여러 문제점들을 가지고 있으며, 특히 은어와 같이 표준화 되지 않은 언어들로 인하여 같은 문장이 시기와 환경에 따라 다른 의미로 해석되는 것을 인지할 수 있어야 한다[2]. 이러한 자연어를 적절하게 처리할 수 있는 자연어 처리 모델의 개발을 위해선 모델의 주기적인 업데이트와 실시간적 데이터의 확보가 중요하다.

III. The Proposed Scheme

본 시스템은 유튜브에서 댓글들을 크롤링하여 해당 댓글들의 감성을 긍정과 부정으로 분류한다. 또한 이렇게 입력받은 자연어 속 단어들의 빈도 수를 합산하여 시각화한다.

유튜브는 다양한 주제에 따른 영상들이 게시된다. 감성을 분석하고자 하는 영상의 주소를 입력 받아 파이썬의 Beautiful Soup4 및 Selenium 라이브러리를 활용하여 자연어 데이터를 크롤링하였다.

감성의 이진분류(긍정, 부정)를 수행하기 위해선 최초 학습과 테스트를 위한 막대한 양의 라벨링 데이터셋을 필요로 하는데, 이는 네이버에서 제공하는 영화 리뷰 데이터셋을 활용하였다[3].

Table 1. 네이버 영화리뷰 라벨링 데이터 테이블

index	comment	label

모델의 생성 및 학습은 파이썬의 Tensorflow 라이브러리를 활용하였으며, 구체적인 학습 기법으로는 딥러닝의 순환신경망인 RNN(Recurrent Neural Network)를 활용하였다[4]. 출력층 레이어의 활성화 함수는 Sigmoid 함수를 활용하였으며, 과적합을 방지하기 위하여 테스트 손실률이 20회 연속 감소하지 않을 경우 학습을 조기 종료하도록 하였다.

이렇게 개발된 모델을 통하여 입력된 문장의 감성을 긍정과 분류로 구분하였다. 빈도수를 고려한 시각화를 위하여 파이썬의 wordCloud 라이브러리를 활용하였다.

```
1/1 [==============================] - 0s 24ms/step
91.89% 확률로 긍정 리뷰입니다.

1/1 [==============================] - 0s 30ms/step
93.68% 확률로 부정 리뷰입니다.
```

Fig. 1. 모델을 활용하여 감성을 분석한 모습

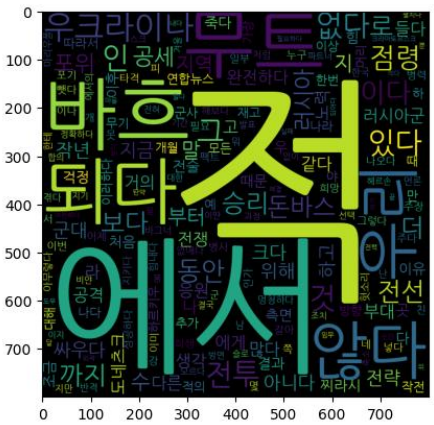

Fig. 2. 등장 빈도 수를 시각화한 모습

IV. Conclusions

본 연구에서 개발한 자연어 분석 시스템은 아직 이진적인 감성 분석, 키워드 추출 등 간단한 기능밖에 지원하지 않는다. 이를 통하여 의미있는 정보를 추출해내는 것은 매우 제한적이며, 추후 연구를 통하여 보다 더 의미있는 데이터를 확보할 수 있도록 하는 기능의 추가가 필요하다. 궁극적으로 하나의 프로그램에서 자연어를 포함한 다양한 데이터를 분석하고, 정보를 추출해낼 수 있는 종합 데이터 관리 프로그램의 개발을 목표로한다.

REFERENCES

[1] Heo, "The Future of AI with the advent of 'ChatGPT', e-patentnews, March, 28, 2023 https://www.e-patentnews.com/9393

[2] Lim Dong-chan. "Study on the Status and Consciousness of Youth's Use of Slang and Slang." Master's thesis in Korea Ajou University, 2015.

[3] https://github.com/e9t/nsmc

[4] Lee Jong-hyuk. "A Comparative Study on the Stock Price Prediction Model Using LSTM and Two-Way Circular Neural Network." Master's thesis in Korea, Seoul National University of Science and Technology, 2019.