# 자연어 처리 모델을 이용한 악플 필터링 시스템 구축

- La belle epoque
황지상, 강범석, 강승범, 김보경, 김인용, 김지영, 노규명, 임호태, 조하담, 하은혜

## 1. 서론

지난 2019년, 가수 구하라와 설리가 극단적인 선택으로 죽음을 선택했습니다. 자살의 이유는 악성 댓글, 줄여서 '악플'로 인한 스트레스 때문이었습니다. 이 두 연예인의 자살은 '네이버 연예' 페이지의 댓글 서비스 폐지로 이어졌습니다. 하지만, 악플로 인한 피해는 그것으로 끝나지 않았습니다. 이제는 연예인뿐 아니라 스포츠 스타, 유튜버, 기자 등 다양한 직군의 사람들이 악플로 인해 고통을 호소하고 있습니다.

지난 8월 3일, 경찰청에 따르면 악플 관련 형사사건은 꾸준히 증가했습니다. 특히 지난해의 경우, 1만 6,633건으로 역대 최다를 기록[1]했습니다. 악플 피해는 유명인들에게만 해당하는 것이 아니었습니다. 판결문을 분석한 결과, 악플로 인해 피해받은 사람의 80%는 다름 아닌 일반인[2]이었습니다. 악플은 더는 특정 집단의 문제가 아닌 사회 전체의 문제로 드러났습니다.

저희 팀은 이번 '2020 국어 정보 처리 시스템 경진대회의 지정 분야'에서 **감성 말뭉치를 분석해 악플을 필터링해주는 시스템을 만들 예정**입니다. 악플 필터링을 통해 악플로 인해 고통을 호소하는 사람들을 줄이겠습니다.

## 2. 관련 연구

### 2.1 감성 분석 연구

자연어에서 감정을 분석하기 위한 시도는 이전부터 있었습니다. 특히 감성 분석 연구는 자연어 처리의 주제로 연구되기 시작했습니다. 특히 경영관리 분야의 학계 및 산업에서 활발히 연구되고 있습니다. 고객이 서비스 혹은 상품에 대해 어떤 생각하는지를 파악해 이윤을 극대화하고자 하는 동기 때문이었습니다.[3] 감정은 크게 이진 분류와 다중 분류를 통해 유형을 구분할 수 있습니다. 이 중 이진 분류의 경우, 영화 평가와 제품 평가에 사용됩니다. **본 프로젝트에서는 이진 분류를 통해 감정을 긍정(Positive)과 부정(Negative)으로 분류**했습니다.

## 3. 프로젝트 설계 및 진행 방법

### 3.1 데이터 수집

본 프로젝트에서는 지도학습 방법을 통해 감성 분석 모델을 구축했습니다. 사용자들의 다양한 분야의 댓글을 수집하기 위해 '네이버 스포츠'의 영상과 '네이버 TV'의 연예 프로그램 Talk을 크롤링했습니다. 여기에 본 공모전에서 제공한 '네이버 영화' 댓글을 사용했습니다.

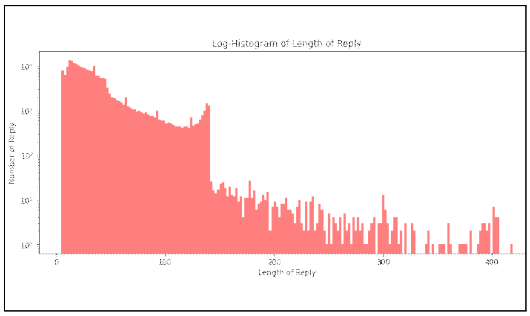
'네이버 스포츠' 의 '많이 본 영상' 랭킹 30위에 달린 댓글들 가운데 2015년부터 2020년 8월 31일까지 6년 치의 데이터를 수집했습니다.

'네이버 TV'의 연예 프로그램 Talk에서는 구독 수가 10만 건 이상, 댓글이 30만 건 이상의 프로그램을 대상으로 했습니다. 총 10개의 프로그램을 선정해 프로그램별 10만

1) 김성호, **"하루 고소 45건, 악플 피해 급증에도 수사는 '뒷전'?"**, 파이낸셜 뉴스, 2020.08.13, https://www.fnnews.com/news/202008031336104724
2) CBS 노컷뉴스, **"'죽음의 악플', 242건 판결문 전수분석해보니..."**, CBS, 2020.02.07, https://www.nocutnews.co.kr/news/5285339
3) Appel, O., F. Chiclana and J. Carter, **"Main concepts, state of the art and future research questions in sentiment analysis", 2015,** Acta Polytechnica Hungarica, Vol.12 No.3, 87~108.

개의 댓글을 크롤링했습니다. 수집 대상으로는 예능 프로그램 '골목식당', '미운 우리 새끼', '1박 2일', '놀면 뭐하니', '나 혼자 산다', '아는 형님'을, 드라마 프로그램 '슬기로운 의사생활', '이태원 클라쓰', '비밀의 숲', '싸이코지만 괜찮아'를 대상으로 했습니다. 이 가운데 5만 개의 데이터를 직접 라벨링해 1차 학습 데이터로 이용했습니다.

이렇게 수집된 댓글은 약 75만 개가 수집됐습니다. 이렇게 수집된 댓글 가운데 약 25만 개를 실험용 데이터와 모델을 구축하기 위한 데이터로 이용했습니다.




<그림 1> 전처리 이전 수집 데이터

<그림 1>은 댓글의 길이입니다. 댓글별 평균 길이는 35.05으로 나타났습니다. 댓글 길이의 중앙값은 26.0입니다.

저희는 저희가 구축한 말뭉치 데이터셋을 연구의 목적으로 공개할 의사가 있음을 밝힙니다.

## 3.2 프로젝트 설계

프로젝트의 진행 순서는 <그림 2>와 같습니다. 3.1에서 언급했던 바와 같이 데이터 수집 단계를 마쳤습니다. 수집된 데이터 중 모델 구축을 위한 라벨링된 데이터는 총 251,081개입니다. 이를 전처리하고 기계학습 기반의 감성 분석 모델을 구축하는 데 사용했습니다. 감성 분석 모델은 기계학습이 기반입니다. 전처리 이후, 데이터의 80%를 학습용 데이터로, 20%의 데이터는 테스트용 데이터로 이용했습니다.


<그림 2> 프로젝트 진행 절차

## 3-3. 데이터 전처리

전체 데이터를 대상으로 전처리 과정을 진행했습니다. 댓글의 길이가 5보다 작은 경우, 감성 분석의 대상에서 제외했습니다. 또한, 댓글의 라벨링을 위해 두 가지 방법을 이용했습니다.

### 3-3-1) 스포츠 댓글

스포츠 댓글의 경우, 해당 글에 달린 '좋아요'와 '싫어요'의 비율을 이용해 라벨링을 실시했습니다. '좋아요'와 '싫어요'를 누른 사용자의 비율 가운데 '좋아요'가 20% 이하이면 '부정', 80% 이상이면 '긍정', 그 외에는 '그레이 에리어'로 보고 0을 라벨링했습니다. 이후 학습 과정에서 0을 다 제외시키고 진행했습니다.

| | 부정 | | 긍정 |
|---|---|---|---|
| 영화 리뷰 | 0 | | 1 |
| | 부정 | 그레이 | 긍정 |
| 스포츠 댓글 | 0.2 | 0.6 | 0.2 |

<표 1> 라벨링 방법

### 3-3-2) 연예 댓글

연예 댓글의 경우, 통계적 기준을 이용할 만한 지표가 뚜렷하지 않았습니다. 저희 팀은 해당 댓글을 직접 라벨링을 실시했습니다. 라벨링의 기준은 다음과 같습니다.

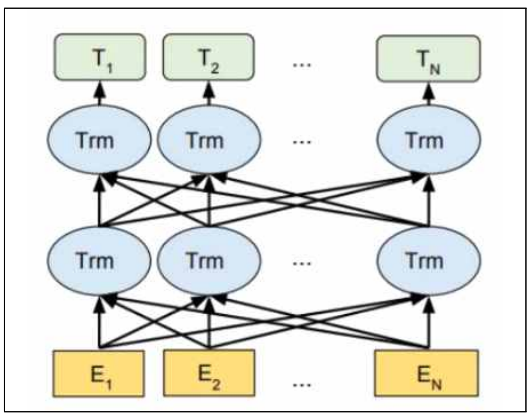
'출연진, 방송 프로그램, 제작진'과 같은 뚜렷한 대상에 대한 칭찬, 응원, 긍정적인 문구에 긍정을 라벨링했습니다. 대상에 대한 비방, 욕설, 실망, 지루함, 짜증, 부정적인 내용에 대해서는 부정을 라벨링했습니다. 또한, 네티즌 간에 싸우는 내용 등 역시 부정으로 라벨링했습니다. 이 외에는 중립을 라

벨링했습니다. 슬픔, 놀람, 아쉬움, 요구, 기다림, 또 대상이 없는 댓글에 중립을 부여했습니다. 이후, 학습 과정에서는 중립에 해당하는 데이터들을 제외했습니다. 긍정, 부정으로 라벨링된 데이터들만 학습에 이용했습니다.

여기에 댓글들의 '악플' 여부를 구분 짓기 위해 악플인지를 판단하는 라벨링을 진행했습니다. 특정 대상에 대한 비판이 넘어서는 댓글을 악플로 판단해 1을 부여하고 아닌 경우에는 0을 부여했습니다. 예를 들어, "방송 너무 못한다."라는 댓글은 비판으로 판단해 0을, "방송 XX 못하네. 꼴 보기 싫다."와 같이 욕설이 섞인 비난은 1을 부여했습니다.

### 3.4 모델 구축 과정

감성 분석을 위해 자연어 처리 분류 모델인 BERT를 이용했습니다. BERT는 'Bidirectional Encoder Representations form Transfomer'의 준말로, "Attention is all you need"에서 소개한 'Transformer'을 활용한 언어 표현 모델입니다. BERT는 사전에 학습된 모델을 목적에 맞게 'Fine-Tuning'해 사용합니다.




<그림 3> BERT 모델 구조

저희 팀은 공개된 한국어 BERT 모델 가운데 KcBERT[4]를 이용해 저희의 데이터셋을 이용해 Fine Tuning을 진행했습니다. 기존에 공개된 KoBert와 같은 한국어 BERT의 경우 정제된 데이터를 기반으로 학습됐습니다. 구어체와 신조어, 오탈자가 많이 등장하는 댓글들과 기존의 한국어 BERT 모델은 적합하지 않다고 판단했습니다. 스포츠와 연예 프로그램에서 사람들이 많이 사용하는 언어를 추가하고자 했습니다.

KcBERT의 경우 15.4GB의 댓글과 대댓글로 학습이 이루어졌습니다. 저희는 사전학습된 KcBERT를 이용해 감성 분석을 진행했습니다.

## 4. 감성 분석 결과

### 4.1 감성 분석 모델 구축

BERT 모델의 장점은 Hugging Face[5]에서 제공하는 툴을 이용해 손쉽게 모델을 튜닝할 수 있다는 점입니다. 1차 심사에서는 KcBERT와 한국어로 학습되지 않은 albert 모델 가운데 'uncased-albert-v2' 모델의 학습 결과를 비교했습니다.

| KcBERT | | uncased-albert-v2 | |
|---|---|---|---|
| Accuracy | 85.71% | Accuracy | 59.84% |

<표 2> 모델 비교(에폭 4회 기준)

한국어 기반의 KcBERT를 이용했을 경우, 감성분석에 있어서 압도적으로 좋은 성과를 만들어낼 수 있음을 확인했습니다.

### 4.2 감성 분석 모델을 활용한 시스템 제안

저희는 감성 분석 모델을 이용해 악플을 필터링하는 방법으로 두 가지를 생각했습니다.

첫 번째는, 감성 예측 결과가 극단적으로 부정으로 예측되는 것을 악플로 보는 방법론입니다. 긍정과 부정에 대한 감성 예측 과정에서 확률이 부여됩니다. 이 가운데 부정으로 예측되는 확률이 0.9 이상이 되는 경우 악플로 보는 방법이 있습니다.

두 번째는, 악플에 대해 추가적으로 라벨

4) KcBert : Korean comments BERT, "**https://github.com/Beomi/KcBERT**"
5) Hugging Face, "**https://huggingface.co/transformers/**"

링한 데이터를 이용하는 것입니다. 연예 프로그램의 경우 전통적인 자연어 처리의 문제와 같이 저희가 직접 데이터에 악플 여부를 체크했습니다. 이런 전통적인 자연어 처리 방법뿐만 아니라, 스포츠 댓글의 경우 '좋아요' 비율이 10% 이하인 댓글을 악플로 보는 방법도 가능합니다.

저희는 2차 심사에서 두 가지 방법론을 모두 활용해 보고 더 적합한 필터링을 적용할 예정입니다.

## 5. 결론

### 5.1 프로젝트 결론 및 시사점

본 프로젝트는 KcBERT를 이용해 감성 분석의 수준을 끌어올릴 수 있었습니다. 이처럼 한국어 기반의 모델을 이용하는 것이 성능을 높이는 데 있어서 중요하다는 사실을 확인했습니다.

또한, 데이터를 손쉽게 라벨링할 방법에 대해 고민할 필요성을 느꼈습니다. 전통적인 라벨링 방법으로는 대용량의 데이터를 처리할 수 없습니다. 스포츠 댓글의 경우 '좋아요'와 '싫어요'의 비율로 라벨링을 처리할 수 있었습니다. 하지만 연예 프로그램의 경우 라벨링을 소프트웨어에 적용할 기준을 찾지 못했습니다. 그 결과, 저희가 수집한 50만 개의 데이터 가운데 10%인 5만 개 밖에 학습에 이용하지 못했습니다. 저희는 2차 심사에서 50만 개의 데이터를 라벨링할 기준을 만들고자 합니다.

### 5.3 추후 프로젝트 진행 방향

저희는 2차 심사까지 총 세 가지의 방향으로 진행할 예정입니다. 첫 번째, 추가된 데이터를 기반으로 모델의 감성 분석 성능을 높일 계획입니다. 50만 개의 연예 프로그램 데이터를 라벨링할 방안을 설립하고 적용할 예정입니다.

두 번째, 악플을 필터링할 시스템을 구축할 예정입니다. 댓글의 긍·부정을 판별하고 부정의 정도에 따라서 악플인지의 여부를 사용자에게 제공할 예정입니다.

세 번째, 구체적인 크롬 확장 프로그램에 적용할 예정입니다. 크롬 확장 프로그램을 통해 해당 프로그램을 사용하는 사용자의 화면에서 악플이 보이지 않게끔 프로그램을 구축할 예정입니다.

## 6. 참고문헌(References)


**[관련 논문]**

1. Appel, O., F. Chiclana and J. Carter, "**Main concepts, state of the art and future research questions in sentiment analysis**", 2015, Acta Polytechnica Hungarica, Vol.12 No.3, 87~108.
2. Vaswani Ashish, et al, "**Attention is all you need**", 2017. 12, NIPS'17 : Proceedings of the 31st International Conference on Neural Information Processing Systems

**[관련 뉴스]**

1. 김성호, "**하루 고소 45건, 악플 피해 급증에도 수사는 '뒷전'?**", 파이낸셜 뉴스, 2020.08.13, https://www.fnnews.com/news/202008031336104724
2. CBS 노컷뉴스, "**'죽음의 악플', 242건 판결문 전수분석해보니...**", CBS, 2020.02.07, https://www.nocutnews.co.kr/news/5285339

**[활용 패키지]**

1. **Transformer**, "https://huggingface.co/transformers/"
2. **KcBert** : Korean comments BERT, "https://github.com/Beomi/KcBERT"
3. **Scrapy**, "https://docs.scrapy.org/en/latest/index.html"

**[분석 도구]**

1. 언어 : **Python**
2. 개발환경 : **Jupyter Notebook, Visual Studio Code**