SNS 빅데이터의 상대 유사도를 반영한 감성 분석 시스템 설계

김봄○ 류소정 오현주 정윤영 김세진 김윤희
숙명여자대학교 컴퓨터과학부
{ pwxcq96; thwjd703; rophy1310; estr1006; wonder0702; yulan }@sookmyung.ac.kr

A Design of a Sentiment Analysis System using Data Similarity among SNS Big Data

Bom Kim○ Sojeong Ryu Hyunju Oh Yoonyoung Jeong Sejin Kim Yoonhee Kim
Dept. of Computer Science
Sookmyung Women's University

요 약

동영상 매체를 활용한 광고가 늘어나고 이에 대한 반응이 댓글로 빠르게 표현됨으로, 이 데이터를 분석하는 것이 중요하다. 이때 사용되는 인터넷 용어는 비속어, 은어, 줄임말 등 비표준어로 다양하게 표현되므로, 기존의 분석 방법으로는 감성 표현의 강도를 분석하는 것이 불가능하다. 본 연구에서는 인터넷 용어와 표준어와의 유사성을 분석하여 감성표현의 강도를 세밀하게 측정하고 고속처리가 가능한 스파크를 사용하여 빠르고 효율적인 감성분석이 가능한 시스템을 제안한다. 그 결과 함께 사용되는 단어들의 밀접도를 기반으로 반어적 표현 등 문맥 상의 의미를 더 정확히 파악할 수 있게 되었다.

1. 서 론

유튜브를 비롯한 영상 콘텐츠 시장의 폭발적인 성장으로 인해 동영상 매체를 활용한 광고가 늘어나고 있으며, 이에 대한 반응을 댓글을 통하여 빠르게 표현하는 추세이다. SNS의 상호 교류의 특성을 이해하고 사용자의 감성 데이터를 분석하는 것은 중요하다. 이에 따라, 동영상 광고 효과를 분석하기 위해 동영상 댓글을 분석하는 시스템에 대한 연구들이 진행되고 있다[4].

인터넷 용어는 표준어 및 그의 변형인 비표준어가 다양하게 발달하여 유사한 감성을 표현하는 다양한 용어들이 존재한다. 따라서 유사 감성어를 군집화하고 유사도 척도를 적용하여 다양한 감성 분류를 정규화 할 필요가 있다. 단어를 공간에 임베딩하는 Word2vec 기법은 기준어 중심으로 유사 단어가 군집을 형성하도록 하여 감성 척도를 함께 표현할 수 있다.

본 연구에서는 SNS의 비속어, 은어, 줄임말을 포함한 인터넷 용어의 감성을 분석할 때에 유사도를 적용하여 유사도가 높은 용어에 가중치를 부여하여 감성표현의 강도를 세밀하게 측정하고자 한다. 시간적 한계를 극복하기 위해 머신러닝에 스파크를 활용하여 빅 데이터 고속처리를 할 수 있으며, 빠르고 효율적인 감성분석시스템(SASDaS: a Sentiment Analysis System using Data Similarity)을 구축한다.

SASDaS의 주요 기능은 동영상 댓글 데이터를 수집하여 저장하는 크롤러와 머신러닝 을 통해 인터넷 용어를 해석할 때에 빅데이터의 고속 처리를 위해 스파크를 사용하고, 유사도에 따른 가중치를 분석하는 인터넷 용어 분석기, 오피니언 마이닝 알고리즘에 가중치를 적용하는 감성 분석기이다. 시스템을 활용하여 유튜브 광고에 적용한 결과 함께 사용되는 단어들의 밀접도를 기반으로 반어적 표현 등 문맥 상의 의미를 더 정확히 파악할 수 있게 되었다.

본 논문의 구성은 다음과 같다. 2장은 관련 연구, 3장은 SASDaS 시스템의 구조를 설명하고, 4장은 실험 환경 및 분석 결과를 설명하며 5장으로 결론을 맺는다.

2. 관련 연구

[1] 논문은 동영상 광고 효과의 전략적인 측정을 위한 방법으로 광고를 실제로 시청한 소비자들을 대상으로 한 설문 조사가 이용되는데 이는 광고주가 결과를 얻어오는 데에 소요되는 시간이 길다.

[2] 논문은 감성 분석을 위해 품사 태깅이 되지 않는 비속어/은어, 이모티콘 등을 분석하였고, 피실험자들에게 의미 평가를 하도록 하여 이를 표준어로 번역하였는데 이는 자동화된 방법론이 아니다.

[3] 논문은 이모티콘의 단순한 의미해석에서 더 나아가 이를 사용한 문장의 감성을 고려하여 18개의 이모티콘의 긍정, 부정을 분석했지만 분석된 이모티콘의 개수가 적다고 판단된다.

[4] 논문에서 머신러닝과 오피니언 마이닝을 이용한 동영상 광고 효과를 분석하는 시스템을 만들었다. 비속어 사전을 구축할 때, word2vec 알고리즘을 통해 구한 단어들의 유사도에 따른 가중치를 두지 않아 감성을 섬세하게 표현할 수 없었다. 또한, 기계학습에 이용되는 댓글 데이터가 매순간 업데이트되어 양이 방대해지는 것을 고려하지 않아 모델을 학습시키는 시간이 오래 걸려 사용자에게 결과를 도출해 내는 과정에 긴 시간이 소요되었다.

3. SASDaS (a Sentiment Analysis System using Data Similarity)

3.1 시스템 구조

* 이 논문은 2019년도 정부재원(과학기술정보통신부 여대학원생 공학연구팀제 지원사업)으로 과학기술정보통신부, 한국연구재단과 한국여성과학기술인지원센터의 지원을 받아 연구되었습니다.

SASDaS는 에서는 댓글 데이터를 기반으로 광고에 대한 사용자 댓글에 포함된 인터넷 용어의 감성을 빠르게 분석하는 시스템이다. 그림 1은 본 시스템의 구조도를 나타낸다. 사용자로부터 분석하고자 하는 광고의 URL을 입력 받아 데이터 크롤러(Data Crawler)를 통해 댓글 데이터를 자동으로 수집한다. 전처리기(Preprocessor-Filtering)에서는 수집된 데이터에서 사용하지 않는 불용어를 제거하고, 한국어 단문 특성을 반영하고 추가로 이모티콘까지 고려하여 명사를 중심으로 데이터를 전처리 한다. 처리된 데이터는 오피니언 마이닝(Opinion Mining)을 적용하여 동영상에 대한 감성 점수를 계산하는데, 이때 형태소 분석기(Lexical Analyzer)와 머신러닝을 통해 비속어, 은어, 줄임말을 분석하여 사전을 구축하는 감성 사전 구축기(Sentiment Dictionary Builder)를 사용한다. 이를 기반으로 감성분석(Sentiment Analysis)을 진행하게 되는데, 이 과정에서 감성을 세부적으로 분석하기 위한 세부 분석 알고리즘을 추가한다. 분석된 결과는 사용자에게 표 또는 그래프로 전달된다.

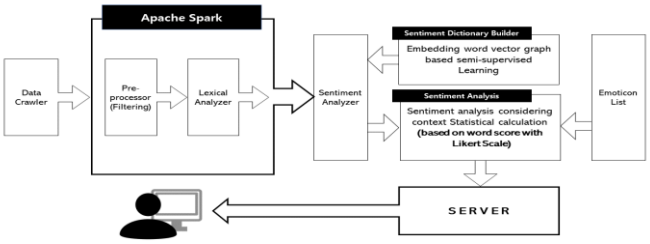

그림 1 본 시스템의 구조도

3.2 스파크를 사용한 인터넷 용어 해석

본 논문에서는 인터넷 상의 댓글들을 기반으로 모델을 학습시킨다. 인터넷 댓글의 특성상 매분 매초마다 생성되기 때문에 그 크기에 제한이 없다. 따라서 시간이 지날수록 데이터의 크기가 증가하기 때문에 그 크기에 맞추어 데이터를 빨리 학습시킬 수 있는 기법이 필요하다.

크롤링한 댓글 데이터를 분산 처리하여 사용하기 위해 Hadoop Distributed File System(HDFS)에 업로드 하였다.

스파크 마스터는 아파치 스파크에서 하둡 클러스터의 전체적인 리소스 관리를 맡고 있는 리소스 매니저인 yarn, serializer 는 스파크에서 제공하는 고성능의 직렬화 라이브러리인 kyro 를 이용한다.

모델을 학습시키기 위해 먼저 스파크 쉘 상에서 트위터가 제공하는 한국어 처리기인 twitter-korean-text 의 정규화 기능을 사용해, 줄 단위의 한국어 텍스트를 처리한다. 이는 오타가 많은 sns 댓글 데이터 특성상 필요하다. 정규화 단계를 완료한 후 문장을 품사 단위로 나누고 이를 문자열로 변환시켜 새로운 Resilient Distributed Data(RDD)를 생성한다. 생성된 RDD 를 word2vec 모델에 학습시켜주면, 스파크는 이를 여러 단계로 나누어 병렬분산 처리를 실행한다.

3.3 유사도를 반영하여 가중치를 적용한 비속어 해석

그림 2는 word2vec에서 코사인 유사도의 예제이다. 'Obama'와 'president'가 밀접한 관계임을 보여주고 있다. 이는 입력한 단어와 의미적으로 유사한 정도를 나타내는 수치이며 이를 가중치로 활용하여 점수를 계산하였다. 이 과정을 통하여 유사한 단어일수록 입력한 단어의 점수에 더 높이 기여하게 되고, 결론적으로 섬세한 단어 의미 분석이 가능해진다.

그림 2 word2vec 코사인 유사도 [8]

다음은 비속어 해석 과정이다.

1) word2vec 모델을 사용하여 학습 데이터에서 유사 단어를 추출하고, 그 중 감성 사전에 등록되어 있는 상위 단어들을 다시 추출한다.
2) 상위 단어의 감성 값을 긍정점수, 중립점수, 부정점수 세 가지로 나누어 각 값에 가중치를 곱한다. 가중치는 word2vec의 유사도 단위인 코사인 유사도를 사용하였다.
3) 각 긍정점수, 중립점수, 부정점수 값에 누적된 점수를 [4] 논문을 기반으로 5점 만점으로 환산한다.

4. 실험 환경 및 결과 분석

본 시스템은 master 서버 1대와 slave 서버 2대로 클러스터를 구성하고 실험을 진행하였다. 실험 진행 환경은 표 1과 같다.

	프로세서	CPU속도	Core 개수	캐시 크기
Jupyter Notebook	Intel Core i5-8400	2.8GHz	헥사 코어	11MB
Spark Master	Intel Core i7-7820X	3.6GHz	옥타 코어	11MB
Spark Slave1	Intel Xeon E5-2630V3	2.4GHz	옥타 코어	20MB
Spark Slave2	Intel Core i7-5820K	3.3GHz	헥사 코어	15MB
	Java	Hadoop	Spark	Scala
소프트웨어 버전	1.8.0_201	2.8.5	2.2.2	2.11.8
	Python			
	3.7.1			

표 1 하드웨어 스펙 및 소프트웨어 리스트

4.1 스파크를 사용한 머신러닝

모델 학습에 사용된 데이터는 126MB의 영화 리뷰[9] 및 유튜브 댓글 텍스트 데이터이다.

	학습 시간	성능 향상도
파이썬	33분	-
Slave 1개	19분	1.65
Slave 2개	13분	2.53

표2 환경에 따른 실험 결과

실험 진행 환경에 따른 결과는 표2와 같다. 이를 통해 파이썬 보다는 스파크 환경에서 실행시간이 단축되었음을 알 수 있으며, 더 나아가 같은 스파크 환경에서는 slave를 추가하였을 때 보다 좋은 성능을 보임을 확인할 수 있었다.

4.2 코사인 유사도를 사용한 가중치 계산

비속어 해석에서 사용한 데이터는 유튜버 '이사배'[7]의 동영상의 댓글을 사용했다. 총 사용한 댓글은 1.27MB이고 여기에 표준어 데이터를 학습시키기 위해 법률개정안 91.7KB를 사용했다.

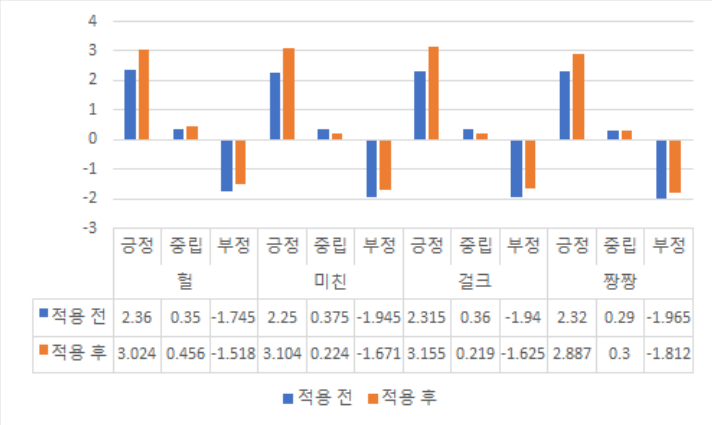

그림 3 비속어 해석 결과 (유사도에 따른 가중치 적용 전/후)

그림 3은 수집한 댓글을 기반으로 비속어인 '헐', '미친', '걸크', '짱짱'의 유사도를 적용하기 전과 적용한 후의 감성 분석 결과이다. '헐'과 '미친'은 인터넷 상에서 긍정에 가까운 감탄사를 나타내는 용어이며, '미친'의 경우, 사전적으로 정의된 부정적인 단어의 뜻이 아닌 긍정의 뜻으로 해석되었다. 또한, '걸크'와 '짱짱' 두 단어 모두 신조어로 '걸크'는 '멋진 여성'을 의미하고 '짱짱'은 긍정적인 감탄사로 두 단어 모두 긍정의 결과로 분류되었다. 하지만 유사도 적용 전의 점수를 보면 네 단어 모두 극 긍정을 표현하는 단어임에도 불구하고 긍정점수와 부정점수의 차이가 크게 다르지 않고 부정의 점수가 크게 나타나는 단점이 있다.

단점 보완을 위해 해당 비속어의 유사어들의 유사도에 따라 가중치를 부여하여 감성 점수를 계산하였다. 논문[4]의 시스템 설계 순서에서 비속어의 감성을 분석하는 부분에 해당한다. 유사도를 적용한 결과는 그림 3에서 유사도 적용 전과 비교하여 극 긍정을 나타내는 단어의 긍정점수가 보완되었다는 것을 알 수 있다.

	헐		미친		걸크		짱짱	
1	미모	0.999200404	ㅠㅜ	0.999325	매일	0.999713	ㅎㅎㅎ	0.999626
2	여신	0.999173888	개	0.999323	매력	0.999675	뷰티	0.999453
3	좋아요	0.999166625	ㅋㅋㅋㅋ	0.999244	색깔	0.999651	매트	0.99939
4	대박	0.998977778	ㅋ	0.999242	분위기	0.999567	약간	0.999385
5	좋다	0.998977661	예뻐	0.999099	ㅋㅋㅋㅋㅋ	0.999541	예전	0.999372

표 5 유사도에 따른 상위 유사어 리스트

표 5는 비속어의 감성을 해석하기 위해 사용한 상위 5개의 유사어와 유사도를 나타낸다. 유사도를 오름차순으로 정렬하여 결과 분석에는 상위 300개를 사용하였으나, 위의 표 3에는 각 비속어당 상위 5개만을 나타냈다. '헐'의 경우에는 '미모', '여신', '좋아요', '대박' 그리고 '좋다'라는 상위 5개의 유사어가 나왔고, 해당 비속어의 감성 점수를 분석하기 위해서 유사도에 따라 가중치를 주었다. 따라서 가장 유사도가 높은 '미모'가 '헐'의 감성 점수를 분석하는데 가장 큰 기여를 하게 된다.

5. 결론

본 논문에서는 SNS 빅데이터의 상대 유사도를 반영한 감성분석 시스템인 SASDaS을 제안하였다. 인터넷 용어인 비속어, 은어 등의 감성을 명확하게 분석하였다. 향후 연구로는 다양한 분야의 댓글을 수집하여 데이터의 감성 분류를 확대하고 댓글 텍스트뿐만 아니라 멀티미디어의 감성 분석에 적용할 계획이다.

참고문헌 (Reference)

[1] 손동진. 2016. 디지털 동영상 플랫폼에 노출되는 동영상광고 크리에이티브의 효과 측정 개념에 대한 연구

[2] 장필식. 2014. 소셜 데이터의 주된 감성분석에 대한 연구

[3] 채인영. 2017. SNS 데이터를 이용한 감성분석 기반의 장소 선호도 분석기법 연구 : 서울시 테마공원을 대상으로

[4] 김세진, 김지은, 성원영, 김봄, 류소정, 이은아, 이지윤, 정윤영, 김윤희. 2018. Youtube 동영상 광고 댓글을 통한 광고 효과 분석 서비스 설계

[5] Gensim, https://radimrehurek.com/gensim/

[6] KoreanSentimentAnalyzer에서 제공하는 감성사전, https://github.com/mrlee23/KoreanSentimentAnalyzer

[7] 이사배유튜브, https://www.youtube.com/channel/UC9kmlDcqksaOnCkC_qzGacA

[8] https://markroxor.github.io/gensim/static/notebooks/WMD_tutorial.html

[9] 영화 리뷰 데이터 https://proinlab.com/archives/1685