WPM(Word Piece Model)을 활용한 구글 플레이스토어 앱의 댓글 감정 분석 연구

박재훈○, 구명완
서강대학교 정보통신대학원
maxuper@sogang.ac.kr, mwkoo@sogang.ac.kr

A Study on the Sentiment analysis of Google Play Store App Comment Based on WPM(Word Piece Model)

Park jae Hoon○, Myong-wan Koo
Sogang University, Graduate School of Information & Technology

요 약

본 논문에서는 한국어 기본 유니트 단위로 WPM을 활용한 구글 플레이 스토어 앱의 댓글 감정분석을 수행하였다. 먼저 자동 띄어쓰기 시스템을 적용한 후, 어절단위, 형태소 분석기, WPM을 각각 적용하여 모델을 생성하고, 로지스틱 회귀(Logistic Regression), 소프트맥스 회귀(Softmax Regression), 서포트 벡터 머신(Support Vector Machine, SVM)등의 알고리즘을 이용하여 댓글 감정(긍정과 부정)을 비교 분석하였다. 그 결과 어절단위, 형태소 분석기보다 WPM이 최대 25%의 향상된 결과를 얻었다. 또한 분류 과정에서 로지스틱회귀, 소프트맥스 회귀보다는 SVM 성능이 우수했으며, SVM의 기본 파라미터({'kernel':('linear'), 'c':[4]})보다 최적의 파라미터를 적용({'kernel': ('linear','rbf', 'sigmoid', 'poly'), 'C':[0.01, 0.1, 1.4.5]} 하였을 때, 최대 91%의 성능이 나타났다.

주제어: WPM, Word Piece Model, 감정 분석, 오피니언 마이닝, 기계 학습

1. 서론

최근 수년간 안드로이드 운영체제 스마트폰이 급속도로 대중화되면서 앱 시장도 끊임없이 성장해 왔다. 사용자들은 앱을 설치하고 사용한 경험을 바탕으로 이에 대한 평가를 앱에 댓글로 표현한다. 또한 사용자들은 자신이 설치 또는 구매 하기 전에 사람들의 반응 즉, 댓글을 이용하여 설치 또는 구매한다. 하지만 앱 댓글은 전체를 읽을 수 없기 때문에 일부 댓글과 댓글 개개의 평점보다는 전체 평점을 참고하여 의사결정을 하는 정도가 대부분이다. 이처럼 전체 평점만을 참고하면 편향적인 습득으로 인하여 이용자가 올바른 정보를 습득한다고 보기 어렵다. 이러한 특성에도 불구하고 댓글은 사용자의 의견을 풍부하게 드러내고 앱을 사용해 보지 않은 다른 이용자들의 선택에 영향을 미친다는 점에서 다양한 실용적 활용성을 갖는 데이터임은 분명하다.

앱 댓글의 '정보' 는 크게 '의견'이라는 두 개의 영역으로 나누어진다. '사실'은 객관적인 정보를 전달하고, '의견'은 사용자의 감정을 표현한다. 앱 댓글은 '의견'을 표현한 정보이다. 댓글의 '의견' 즉 감정과 평가의 측면을 다룬다는 점에서 본 연구는 감성을 분석한다고 할 수 있다. 즉 이 연구는 앱 댓글의 평가와 평점을 기반으로 감성 사전을 구축하고 기계 학습 기반의 알고리즘으로 검증하고 평점을 예측함으로써 수많은 인터넷 문서의 '의견'문서에서 '평가'나 '의견'의 정도를 긍정과 부정으로 변환하여 사용자에게 직관적인 정보를 제공하는 것과 관련된 연구이다.[1]

본 논문에서는 구글 플레이에서 제공되는 앱의 댓글을 수집(2016.07 이전의 카테고리의 앱 댓글 총 1,580,234건 중 무작위로 선정된 682,000건)하여 댓글 감정 분석 결과를 예측하고자 한다. 특히 스마트폰 환경에서 댓글을 쓸 때 나타나는 오타, 띄어쓰기의 부정확함을 줄이기 위해서 띄어쓰기 교정기를 사용하고, 구글의 음성 검색(Voice Search)에 적용된 WPM(Word Piece Model)기법과 형태소 분석기를 사용하여 한국어 기본 유니트 단위를 도출한다. 그리고 기계학습 모델인 로지스틱 회귀(Logistic Regression), 소프트맥스 회귀(Softmax Regression), 서포트 벡터 머신(Support Vector Machine, SVM)을 사용하여 각각 성능 실험 결과를 도출하였다.

2. 관련 연구

2.1 자동 띄어쓰기 시스템

텍스트 분석에서 가장 기초적인 작업은 텍스트로부터 단어를 식별하고 추출하는 토큰화(Tokenization)라고 할 수 있다. 한국어는 '어절(語節)'로 토큰의 단위 기준으로 보고, 중국어나 일본어와 같이 어절 경계 표지가 없는 언어와는 달리, 어절과 어절 사이에 공백을 두어 띄어쓰기를 하도록 규정하였다. 한국어에 있어서 잘못된 띄어쓰기는 중의성(ambiguity)을 유발 시키거나 텍스트 분석에서 잡음(noise)을 일으켜 오히려 토큰화를 방해하며, 가독성을 떨어뜨린다. 이와 같이 한국어에서 띄어쓰

기는 텍스트에 대한 사용자 가독성만큼이나 기계 가독성에도 영향을 주는 중요한 요소이다. 문장 내의 띄어쓰기 오류는 많은 문법적, 의미적 모호성을 일으키며, 때로는 형태소 분석을 불가능하게 만들기도 한다.[2]

본 논문에서는 사용자가 띄어쓰기를 고려하지 않고 댓글을 쓰는 경우가 많다고 가정하고, 자동 띄어쓰기 시스템을 적용한 것과 하지 않은 것을 비교하여 실험하였다.

2.2 WPM(Word Piece Model)

WPM은 2012년 구글에서 제안한 구글 일본과 구글 한국에 적용된 성공적인 음성 검색 시스템 구축을 위한 방법이다. 기존의 자연어 처리에서는 형태소 분석, 통사 분석, 의미 분석, 화용 분석의 4단계로 진행되나, WPM은 음절을 기반으로 유닛을 코드화 시킨 후 사용빈도수에 따라 조합을 하여 새로운 유닛을 생성한다. 즉, WPM은 음절을 기반으로 통계적 기법을 활용하여 사용빈도수가 높은 음절을 합쳐서 사전을 만드는 방법이다.

WPM의 장점은 언어에 독립적이며, 통계적인 방식을 사용하므로 특정 도메인 또는 아직 의미를 파악하지 못한 상태에서도 적용할 수 있다.[3]

본 논문에서는 WPM을 적용한 것과 하지 않은 것을 비교하여 실험하였다.

2.3 로지스틱 회귀 (Logistic Regression)

로지스틱 회귀는 logit function = $y = 1/(1+e^{-x})$ 을 사용한다는 의미의 logistic, 그리고 output과 input사이의 관계를 찾는다는 의미의 regression을 합친 단어이다. 이 알고리즘의 원리는 linear regression과 비슷하게 잘못 분류되는 instance를 최소화하는 선을 그리고 그 선의 parameter를 구하는 것을 말한다.[4]

본 논문에서는 자동 띄어쓰기 시스템과 WPM을 적용한 데이터를 기본적인 분류기로 로지스틱 회귀를 적용하여 실험하였다.

2.4 소프트맥스 회귀 (Softmax Regression)

소프트맥스 회귀는 로지스틱 회귀의 Multiclass 버전이다. 모든 출력값의 분모의 총합을 1로 정규화 시킨다. 그리고, 각각의 출력별로 비율을 책정한다. 총합은 계속 1로 만들게 되므로 한 개의 강력한 feature가 나타나면 이 값은 1로 수렴을 하는 과정에서 나머지 값에도 영향이 미쳐서 0으로 수렴하게 만들어 학습의 가속화가 생긴다. 본 논문에서는 output의 성능이 기존의 시그모이드 함수보다 소프트맥스가 성능이 좋다고 알려져 있어서 적용해 보았다.

2.5 서포트 벡터 머신 (Support Vector Machine, SVM)

서포트 벡터 머신(SVM)은 최초에 러시아의 Vladimir Vapnik 에 의해 1979년에 제안 되었으나 그 당시에는 크게 주목 받지 못 하였다. 그러나 그 이후에 기계학습 분야에서 상대적으로 뛰어난 성능이 확인되어 많이 응용되고 있다. 또한 서포트 벡터 머신 (SVM)은 분류 문제를 해결 할 때 자체적으로 복잡도 조정이 되어 과적합을 방지하는 특징이 있고, 상대적으로 사용하기도 쉽고 약간의 튜닝으로도 다양한 문제에 적용할 수 있다[5].

본 논문에서는 SVM이 속도는 느리지만, 분류기로서 성능이 좋아서 옵션 튜닝을 하면서 실험해보았다.

3. 실험데이터의 구성

3.1 수집 Set의 준비

본 실험에서 사용한 데이터는 구글 플레이 스토어에서 수집하였다. 앱이 중복될 수 있기에 추천 앱 페이지는 수집하지 않았으며, 26개 카테고리에서 300위까지의 앱을 수집한 결과 앱의 개수는 약 7800개이다. 다시 앱마다 댓글을 수집하였는데, 댓글의 개수는 약 312,000 이다. 수집된 데이터는 데이터 개수마다 정확률 분포를 확인하기 위해 표1처럼 1000개부터 1000개 단위로 증가하여 20,000개까지 20단계로 나누어 수집Set을 만들었다.

데이터 개수 내에 댓글 문장은 중복이 되지 않도록 분리했으며, 정확률에 대한 객관성을 유지하기 위해 10-fold cross validation을 적용하였다.

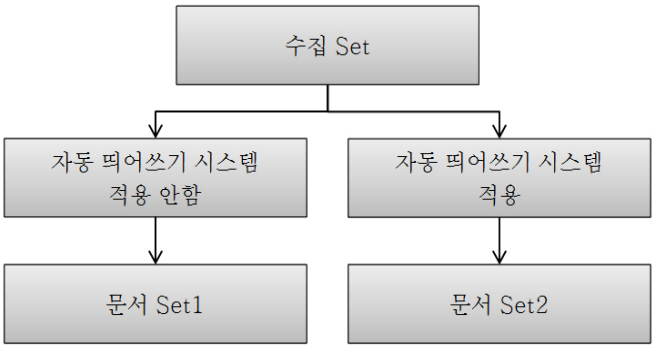
표1. 실험 데이터의 구성

단계	데이터 개수 (긍정/부정)	훈련데이터 (긍정/부정)	테스트데이터 (긍정/부정)
1	1000/1000	900/900	100/100
2	2000/2000	1800/1800	200/200
3	3000/3000	2700/2700	300/300
4	4000/4000	3600/3600	400/400
5	5000/5000	4500/4500	500/500
6	6000/6000	5400/5400	600/600
7	7000/7000	6300/6300	700/700
8	8000/8000	7200/7200	800/800
9	9000/9000	8100/8100	900/900
10	10000/10000	9000/9000	1000/1000
11	11000/11000	9900/9900	1100/1100
12	12000/12000	10800/10800	1200/1200
13	13000/13000	11700/11700	1300/1300
14	14000/14000	12600/12600	1400/1400
15	15000/15000	13500/13500	1500/1500
16	16000/16000	14400/14400	1600/1600
17	17000/17000	15300/15300	1700/1700
18	18000/18000	16200/16200	1800/1800
19	19000/19000	17100/17100	1900/1900
20	20000/20000	18000/18000	2000/2000

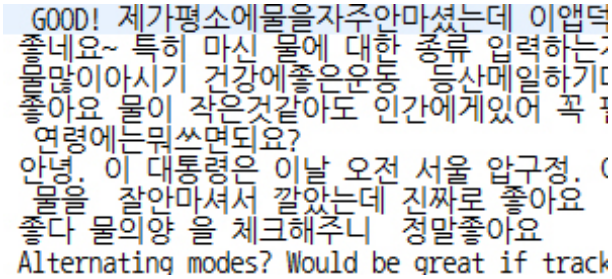
3.2 자동 띄어쓰기 시스템의 적용

수집Set을 그림1과 같이 자동 띄어쓰기 시스템을 적용하지 않은 문서 Set1과 적용한 문서 Set2를 생성하였다.

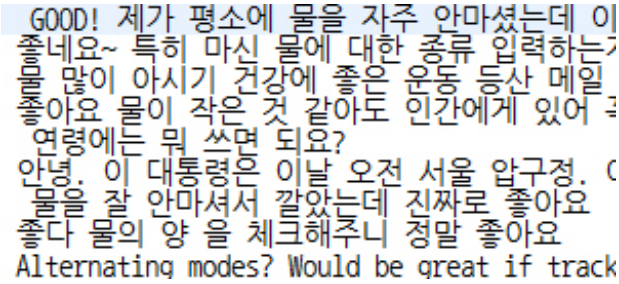
문서 Set1은 자동 띄어쓰기 시스템의 적용하지 않았기 때문에 수집 Set과 구조와 내용이 같다. 그림3의 문서 Set2의 댓글 내용과 그림2의 문서 Set1의 댓글 내용을 비교해보면 띄어쓰기가 되어 있음을 알 수 있다.



[그림 1] 자동 띄어쓰기 시스템의 적용


GOOD! 제가평소에물을자주안마셨는데 이앱덕
좋네요~ 특히 마신 물에 대한 종류 입력하는
물많이아시기 건강에좋은운동 등산메일하기
좋아요 물이 작은것같아도 인간에게있어 꼭
연령에는뭐쓰면되요?
안녕. 이 대통령은 이날 오전 서울 압구정.
물을 잘안마셔서 깔았는데 진짜로 좋아요
좋다 물의양 을 체크해주니 정말좋아요
Alternating modes? Would be great if track

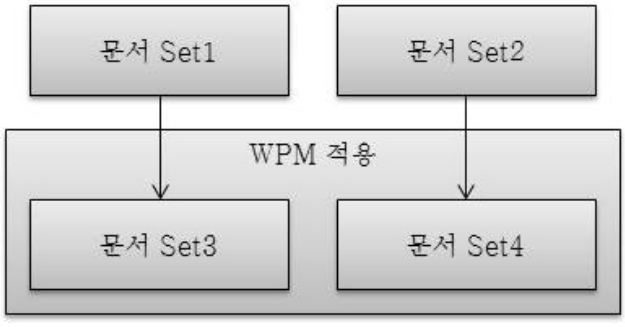
[그림 2] 수집 Set과 같은 문서 Set1의 내용 일부


GOOD! 제가 평소에 물을 자주 안마셨는데 이
좋네요~ 특히 마신 물에 대한 종류 입력하는
물 많이 아시기 건강에 좋은 운동 등산 메일
좋아요 물이 작은 것 같아도 인간에게 있어
연령에는 뭐 쓰면 되요?
안녕. 이 대통령은 이날 오전 서울 압구정.
물을 잘 안마셔서 깔았는데 진짜로 좋아요
좋다 물의 양 을 체크해주니 정말 좋아요
Alternating modes? Would be great if track

[그림 3] 자동 띄어쓰기 시스템을 적용한 문서 Set2의 내용 일부

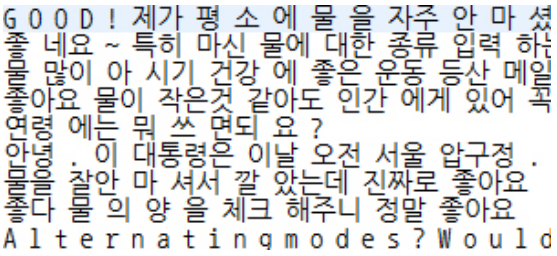
3.3 WPM의 적용

문서 Set1과 문서 Set2는 어절단위와 형태소 분석기용 문서 Set이다. 본 논문은 WPM을 활용한 실험이기 때문에 그림 4와 같이 문서 Set3과 문서 Set4를 생성하였다.

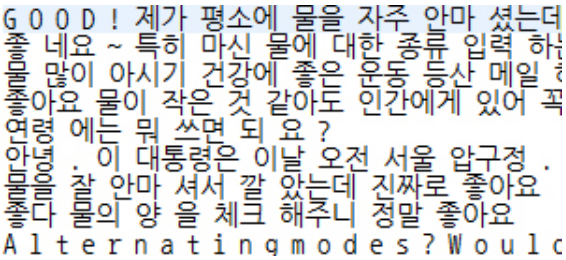


[그림 4] WPM이 적용된 문서 Set3과 문서 Set4

문서 Set3의 댓글 내용은 그림 5와 같이 수집Set에서 자동띄어쓰기를 적용하지 않고, WPM만을 적용한 것이고, 문서 Set4의 댓글 내용은 그림 6과 같이 수집Set에서 자동 띄어쓰기를 적용한 후 WPM을 적용한 것이다.


GOOD! 제가 평 소 에 물 을 자주 안 마 셨
좋 네요 ~ 특히 마신 물에 대한 종류 입력 하
물 많이 아 시기 건강 에 좋은 운동 등산 메일
좋아요 물이 작은것 같아도 인간 에게 있어 꼭
연령 에는 뭐 쓰 면되 요 ?
안녕 . 이 대통령은 이날 오전 서울 압구정 .
물을 찰안 마 셔서 깔 았는데 진짜로 좋아요
좋다 물 의 양 을 체크 해주니 정말 좋아요
A l t e r n a t i n g m o d e s ? W o u l d

[그림 5] 문서 Set1에 WPM을 적용한 문서 Set3의 내용 일부


GOOD! 제가 평소에 물을 자주 안마 셨는데
좋 네요 ~ 특히 마신 물에 대한 종류 입력 하
물 많이 아시기 건강에 좋은 운동 등산 메일
좋아요 물이 작은 것 같아도 인간에게 있어 꼭
연령 에는 뭐 쓰면 되 요 ?
안녕 . 이 대통령은 이날 오전 서울 압구정 .
물을 잘 안마 셔서 깔 았는데 진짜로 좋아요
좋다 물의 양 을 체크 해주니 정말 좋아요
A l t e r n a t i n g m o d e s ? W o u l d

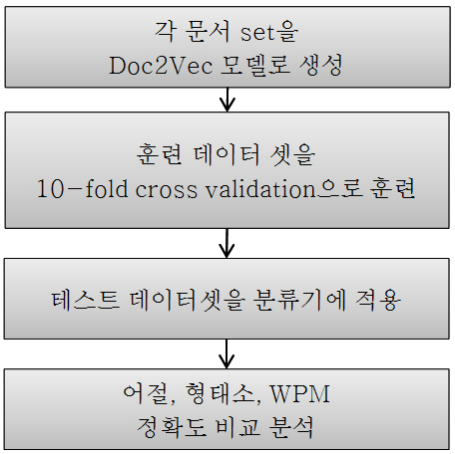
[그림 6] 문서 Set2에 WPM을 적용한 문서 Set4의 내용 일부

4. 실험 및 결과

4.1 데이터의 처리와 분류기(Classifier) 적용 과정

각 문서 set은 Python의 Gensim 라이브러리의 Doc2Vec을 통해 300차원으로 Vector화시켜 모델로 저장하였다.

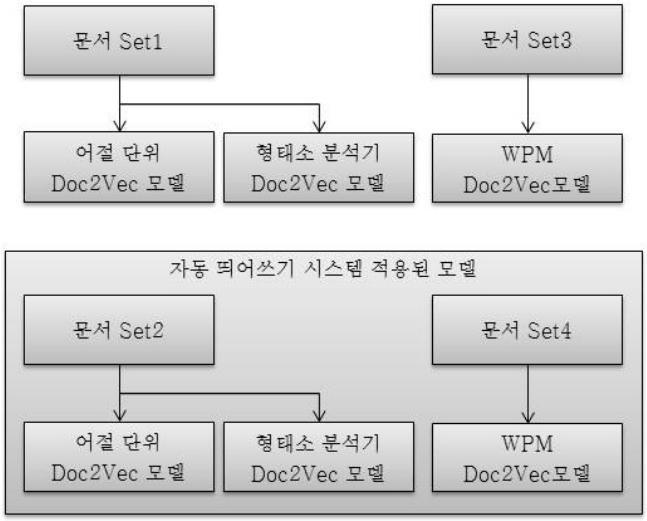
모델을 생성 후에 배열에 저장하여, 훈련데이터와 테스트데이터를 10-fold cross validation을 적용하여 30회 반복 훈련하고, 테스트 데이터셋을 분류기에 적용하였다. 분류기로는 로지스틱 회귀, 소프트맥스 회귀, SVM을 순차 적용하고, 어절, 형태소 분석기, WPM별로 각각 정확률을 산출하여 비교분석 하였다.



[그림 7] 데이터 처리와 분류기 적용 과정

4.2 문서 Set으로 모델 생성

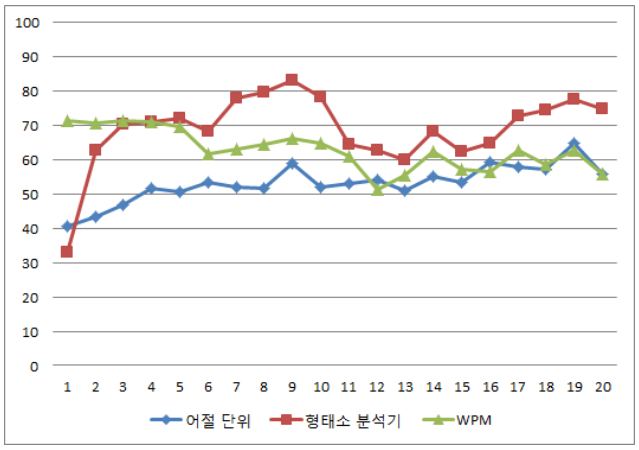
4개의 문서 Set으로 6개의 모델을 생성하였다.



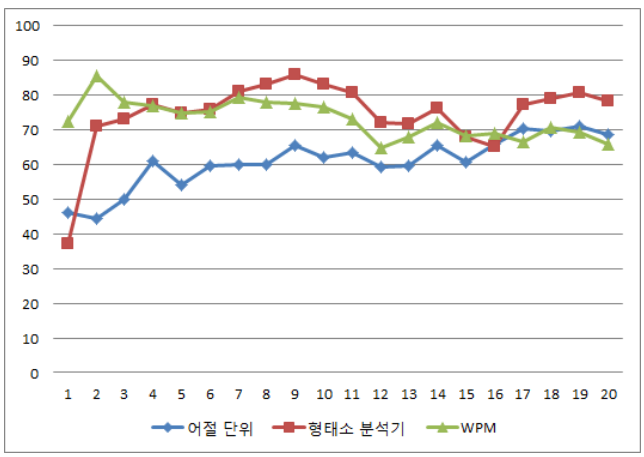
[그림 8] 문서 Set을 모델로 생성

4.3 문서 Set1과 문서 Set3 분류기 실험 결과

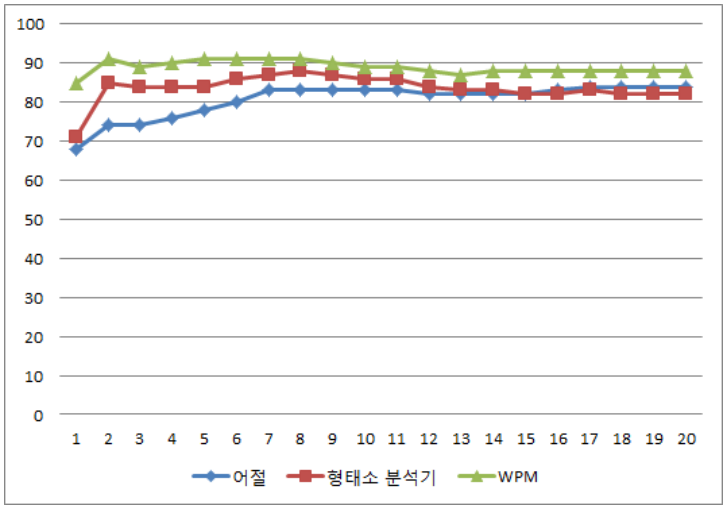
문서 Set1과 문서 Set3의 3가지 모델, 즉 어절 단위, 형태소 분석기, WPM Doc2Vec 모델을 분류기에 적용한 결과는 다음과 같다.



[그림 9] 로지스틱 회귀의 단계별 정확률 결과



[그림 10] 소프트맥스 회귀의 단계별 정확률 결과



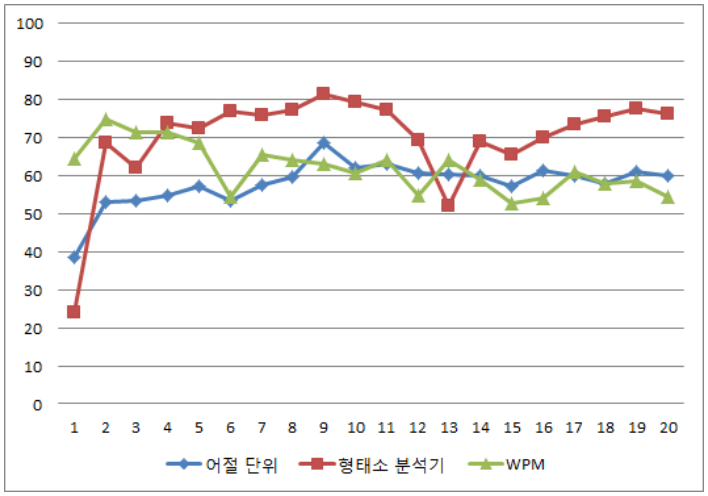
[그림 11] SVM의 단계별 정확률 결과

그림 9, 그림 10을 보면 단계가 증가할수록 성능이 나아지고 있다. 그런데, 그림 11에서처럼 SVM 분류기는 어절, 형태소 분석기, WPM 모두 다른 분류기보다 성능이 고르고 우수하게 나타났다. 특히 WPM과 SVM 분류기 적용 시 2단계(2000/2000)부터 10단계(10,000/10,000)사이에 최고 91%의 성능을 보여주었다.

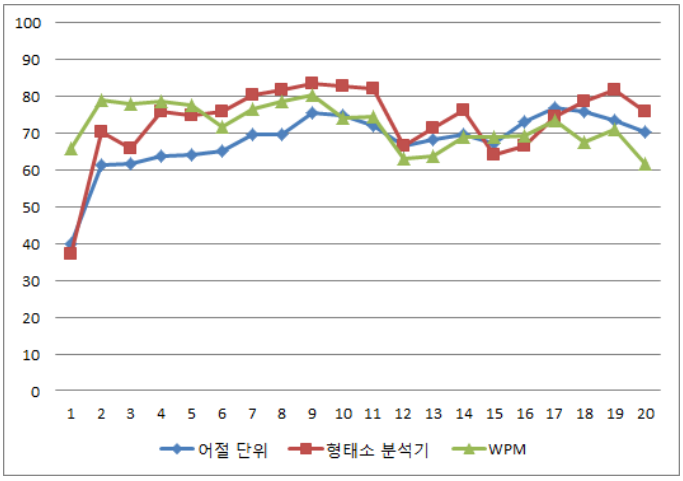
성능 향상은 1단계에서 형태소 분석기와 로지스틱 회귀 분류기의 성능이 33%이고, WPM과 SVM분류기는 85% 성능을 보여 약157% 개선되었음을 알 수 있다. 훈련데이터가 8000개인 8단계에서도 어절단위와 로지스틱 분류기의 성능은 51.74% 이고, WPM과 SVM분류기는 91%를 보여 약 76%의 성능향상을 보여주었다.

4.4 문서 Set2와 문서 Set4 분류기 실험 결과

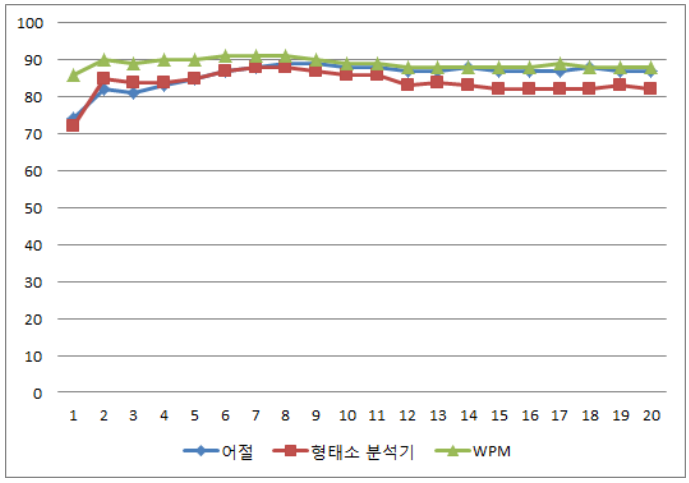
자동 띄어쓰기 시스템을 적용한 문서 Set2와 문서 Set4의 어절 단위, 형태소 분석기, WPM Doc2Vec 모델을 분류기에 적용한 결과는 다음과 같다.



[그림 12] 로지스틱 회귀의 단계별 정확률 결과



[그림 13] 소프트맥스 회귀의 단계별 정확률 결과



[그림 14] SVM의 단계별 정확률 결과

그림 12, 그림 13, 그림 14는 자동 띄어쓰기 시스템을 적용하지 않은 Doc2Vec 모델보다 전반적으로 더 나은 성능을 보여주었다.

6가지의 실험을 비교해 보면 WPM은 1단계부터 85%의 성능을 보여주었지만, 형태소 분석기는 1단계에서는 최저 24%, 2단계에서 68.75%의 성능을 보여주었기에 최소 2000개 이상의 데이터가 확보되어야 할 것으로 보인다.

분류기에서는 로지스틱 회귀보다 소프트맥스 회귀, SVM 순으로 나은 성능을 보이고 있다. 하지만, 댓글의 내용이 부정확한 점을 감안하더라도 SVM은 모든 부분에서 고르고, 높은 성능을 보여주었다.

5. 결론

본 논문에서는 구글 플레이 스토어 앱의 댓글 감정 분석을 WPM의 활용을 했을 때의 성능을 측정하는 것이 목적이었다. 실험을 통해 동일한 분류기(Classifier)에서도 WPM이 우수한 성능을 보여주고 있으며, 성능이 가장 좋은 SVM 분류기에서도 WPM을 적용한 결과가 어절단위나, 형태소 분석기보다 최대 25% 개선이 되었음을 보여주고 있다. 또한, 자동띄어쓰기 시스템을 적용하더라도 어절단위보다 최대 16.2% 만큼 향상됨을 알 수 있었으며, 표 2 에서처럼 자동 띄어쓰기 시스템의 적용 유무와 관계없이 최대 91%로 동일한 결과를 나타냈으며, 데이터 개수에 따른 분포에서도 비슷한 수치를 보여고 있다.

표2. SVM분류기에서 자동 띄어쓰기 적용에 따른 WPM비교

단계	자동 띄어쓰기 적용안함	자동 띄어쓰기 적용함
1	85 %	86 %
2	91 %	90 %
3	89 %	89 %
4	90 %	90 %
5	91 %	90 %
6	91 %	91 %
7	91 %	91 %
8	91 %	91 %
9	90 %	90 %
10	89 %	89 %
11	89 %	89 %
12	88 %	88 %
13	87 %	88 %
14	88 %	88 %
15	88 %	88 %
16	88 %	88 %
17	88 %	89 %
18	88 %	88 %
19	88 %	88 %
20	88 %	88 %

WPM은 훈련 데이터가 적더라도 최소 2,000개만 확보된다면 어절단위나 형태소 분석기보다도 우수한 성능을 보여주고 있다. 향후, 기계 학습의 최신 기술인 ANN(Artificial Neural Network), CNN(Convolutional Neural Network) 등을 도입하여 분류기의 성능을 개선한다면 정확률은 더 개선될 것으로 예상된다.

참고문헌

[1] 조정태, “영화 리뷰 감성 분석을 통한 평점 예측 연구”, 충북대학교 대학원 석사논문, 2015.

[2] 강미영, 정성원, 권혁철, “어절 내의 형태소 범주 패턴에 기반한 통계적 자동 띄어쓰기 시스템”, 정보과학논문지, 2006.

[3] Mike Schuster and Kaisuke Nakajima, “JAPANESE AND KOREAN VOICE SEARCH”, Google Inc, USA, 2012.

[4] 곽동민, 박세원, 이한남, “Machine Learning to Deep Learning", 딥큐먼 인공지능 리서치그룹, 2015.

[5] 박종일, “시계열 데이터 전처리와 특징 선택을 활용한 기계 학습 기반의 효율적인 주가 방향성 예측 모델에 관한 연구”, 서강대 정보통신대학원 석사논문, 2016.