딥러닝 기반 리뷰 유사도 검사를 통한 온라인 쇼핑몰의 가짜 리뷰 탐지 성능 분석

Performance Analysis of Fake Review Detection of Online Shopping Mall through Deep Learning Based Review Similarity Test

저자 (Authors)
김가연, 정혜리, 민지수, 김혁만, 이경용
Gayeon Kim, Hyeri Jung, Jisoo Min, Hyeokman Kim, Kyungyong Lee

출처 (Source)
한국정보과학회 학술발표논문집 , 2018.12, 121-123(3 pages)

발행처 (Publisher)
한국정보과학회
KOREA INFORMATION SCIENCE SOCIETY

URL
http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE07613523

APA Style
김가연, 정혜리, 민지수, 김혁만, 이경용 (2018). 딥러닝 기반 리뷰 유사도 검사를 통한 온라인 쇼핑몰의 가짜 리뷰 탐지 성능 분석. 한국정보과학회 학술발표논문집, 121-123

딥러닝 기반 리뷰 유사도 검사를 통한 온라인 쇼핑몰의 가짜 리뷰 탐지 성능 분석

김가연○, 정혜리, 민지수, 김혁만, 이경용
국민대학교 컴퓨터공학부
ggyyk137@gmail.com, kbi09@kookmin.ac.kr, jsmin0415@gmail.com, hmkim@kookmin.ac.kr, leeky@kookmin.ac.kr

Performance Analysis of Fake Review Detection of Online Shopping Mall through Deep Learning Based Review Similarity Test

Gayeon Kim○, Hyeri Jung, Jisoo Min, Hyeokman Kim, Kyungyong Lee
Department of Computer Science, Kookmin University, Seoul, South Korea

요 약

온라인 쇼핑몰에 의도적으로 긍정적인 가짜 리뷰를 남기는 스패머(spammer)들이 생겨나면서 이들을 탐지하기 위해 스패머의 리뷰 내용이 유사하다는 점을 이용하고 있다. 리뷰 내용의 유사도를 분석하는 방법으로 여러 알고리즘이 사용된다. 그 중 Doc2Vec은 문서에서 텍스트의 의미를 추출하는 알고리즘으로 스팸 리뷰 탐지에서는 리뷰 내용의 유사성을 확인하는 데 쓰인다. 하지만 스패머들은 기계적으로 동일한 내용의 리뷰를 작성하는 경향이 있으므로 리뷰 내용의 의미를 추출하는 Doc2Vec이 Term-frequency나 TF-IDF와 같은 단어의 개수를 세어 벡터화하는 알고리즘보다 스팸 리뷰어 탐지에 있어 낮은 성능을 보인다. 따라서 본 논문에서는 동일한 리뷰를 작성하는 스팸 리뷰어들의 패턴에 Doc2Vec의 리뷰 내용 유사도 분석이 적절하지 않다는 것을 검증하고자 한다.

1. 서 론

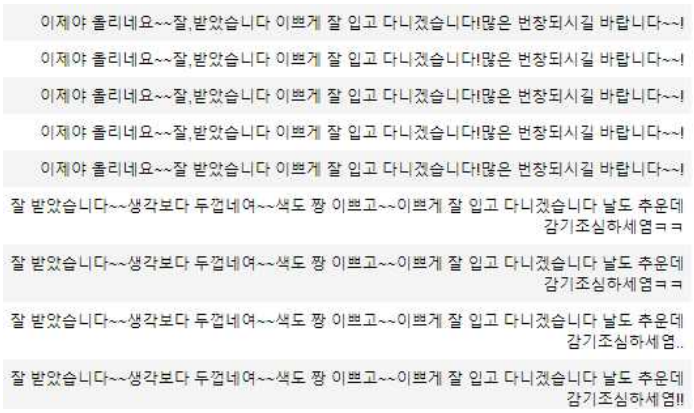
온라인 쇼핑몰을 통해 제품을 구매하는 소비자들이 증가하면서 판매증진을 위한 가짜 리뷰들이 나타나고 있다. 판매자들은 자신의 사이트에 긍정적인 가짜 리뷰들을 기재하고 소비자들에게 거짓된 정보를 제공한다. 이에 따라, 가짜 리뷰 탐지에 대한 필요성이 대두되고 있다[1].

가짜 리뷰의 대표적인 특징으로는 내용의 유사성이 있으며 이와 관련된 알고리즘들을 이용하여 가짜 리뷰들을 탐지하는 연구가 진행되고 있다[2]. Doc2Vec은 문서 검색 시 텍스트의 의미를 추출하는 알고리즘으로 리뷰 내용의 유사성을 찾는다. 하지만 대부분의 스팸 리뷰어(스패머)들은 단순히 내용을 복사하는 수준의 가짜 리뷰를 작성하고 있으므로 내용의 유사도로 가짜 리뷰를 판단하는 Doc2Vec은 다른 알고리즘에 비해 좋은 성능을 발휘하지 못한다.

본 논문에서는 Doc2Vec이 리뷰 의미의 유사도를 통해 가짜 리뷰를 찾는다는 점이 스패머들의 리뷰 작성 패턴과 맞지 않는다는 점을 검증하고자 한다. 실험 데이터에 스패머와 논스패머에 대한 정보가 없으므로 스패머들의 특징을 이용하여 탐지한 스패머들[3]을 ground truth로 설정한다. 이후, Term-Frequency(TF), TF-IDF, Doc2Vec 세 알고리즘으로 탐지한 스패머들과 비교하여 Precision과 Recall로 결과를 분석한다.

2. 본 론

2.1 가짜 리뷰 현황



[그림 1] 스패머의 리뷰

[그림 1]에서와 같이 스패머들은 리뷰를 복사-붙여넣기 방식으로 작성하는 경향이 있다. [그림 1]의 스패머는 197개의 리뷰를 작성하였으며 그중 약 170개의 리뷰 내용이 동일했다.

2.2 Doc2Vec

Doc2Vec은 Word2Vec을 확장한 것으로 하나의 문서(뉴

스 기사, 블로그 글 등)를 실수 벡터로 embedding하는 알고리즘이다. Doc2Vec은 단순히 단어의 출현 빈도만을 따지지 않고 각 단어 사이의 거리를 기준으로 벡터화한다[4]. 즉, 단어의 의미를 고려하여 벡터로 만든다.

2.3 실험 및 평가

2.3.1 실험환경

BeautifulSoup 라이브러리를 이용하여 리뷰를 크롤링하였고, python2.7에서 pandas 라이브러리를 이용하여 데이터를 분석하고 처리하였다. 리뷰를 형태소 단위로 토큰화하기 위해 Konlpy의 Twitter클래스[5]를 사용하였다. 이후, 토큰화된 리뷰들을 벡터화하기 위해 세 가지 알고리즘을 적용하였다. TF와 TF-IDF는 각각 sklearn의 CountVectorizer[6]와 TfidfVectorizer[7]를 사용하였고, Doc2Vec은 gensim의 Doc2Vec[8]을 사용하였다.

2.3.2 실험 데이터

실험에는 온라인 쇼핑몰인 A 쇼핑몰과 B 쇼핑몰을 크롤링한 리뷰를 사용하였다. A 쇼핑몰과 B 쇼핑몰은 각각 772,829개, 304,756개의 리뷰가 존재하며, 리뷰 작성자의 id, 리뷰 내용, 리뷰를 작성한 상품의 id, 리뷰 점수 등을 포함한다. A 쇼핑몰은 리뷰 작성 날짜를 추가로 포함한다. 리뷰의 유사도 파악의 정확성을 위해 리뷰를 2개 이하로 작성한 리뷰어들은 제외하고 실험을 진행하였다. 또한, A 쇼핑몰의 경우 리뷰 작성자가 네이버페이로 구매하였을 때 리뷰 작성자의 id가 네이버페이로 표시된다. 이는 사용자별로 리뷰를 모으기에 어려움이 있으므로 제외하고 720,290개의 리뷰로 실험을 진행하였다.

2.3.3 실험방법

본 실험에서는 ground truth가 존재하지 않으므로 스패머들의 특징을 이용하여 탐지한 스패머들을 ground truth로 이용하였다. 실험에서 이용한 스패머들의 특징은 다음과 같다.

[표 1] 스패머의 특징

특징	내용
1	하루 동안 특정 제품에 n_1번 이상 리뷰를 작성한 리뷰어
2	작성한 전체 리뷰수가 n_2개 이상인 리뷰어
3	제품을 판매하는 쇼핑몰의 상호명을 n_3번 이상 언급한 리뷰어
4	작성한 전체 리뷰의 평균 점수가 n_4이상인 리뷰어
5	한 제품에 n_5번 이상 리뷰를 남긴 리뷰어
6	한 날짜에 n_6번 이상 리뷰를 남긴 리뷰어

위 6가지의 특징을 모두 만족하는 리뷰어는 확실한 스패머라고 간주하고 6가지 중 하나라도 만족한다면 스패머로 의심한다. 본 실험에서는 의심되는 스패머의 n_1~n_6 값을 각각 30, 100, 15, 5.0, 10, 30으로 설정하였다. 위 값들은 여러 번의 실험을 통해 적정값으로 설정한 것이다. 이때, B 쇼핑몰은 리뷰 작성 날짜 정보를 포함하지 않으므로 위 특징 중 4가지만 사용하였다.

각각의 리뷰어들이 작성한 리뷰를 형태소별로 토큰화하여 TF, TF-IDF, Doc2Vec으로 벡터화한다. 이후, 한 리뷰어에 존재하는 벡터들 간의 코사인 유사도를 구하여 리뷰간의 유사도를 파악한다. 한 리뷰어가 n개의 리뷰를 작성했다면 $\frac{n(n-1)}{2}$개의 코사인 유사도가 계산된다. 리뷰어마다 코사인 유사도의 최댓값, 상위 25% 값, 중간값, 상위 75% 값을 구한다. 스패머는 이 4개의 값이 전체 리뷰어의 최댓값, 상위 25% 값, 중간값, 상위 75% 값에 대해 각각 상위 25%에 해당하는 리뷰어들을 의미한다. 이때, 알고리즘별로 탐지한 스패머의 수가 확연히 다른 경우 전체 리뷰어의 상위 25% 이내의 값으로 각각 조절하여 세 알고리즘 중 탐지한 스패머의 수가 가장 적은 쪽으로 스패머의 수를 맞춰준다.

2.3.4 평가도구

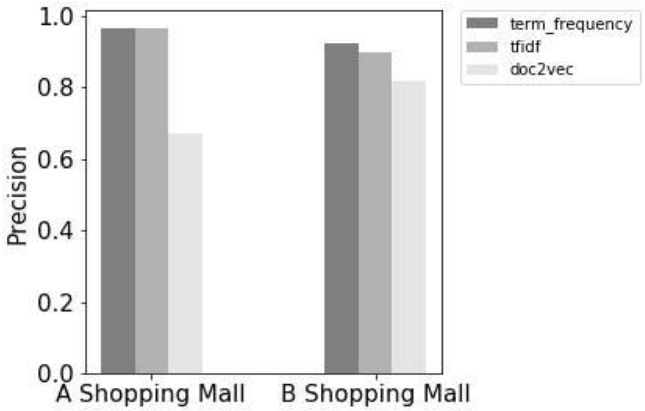
[표 2] Precision과 Recall

	실제 스패머	실제 논스패머
예측한 스패머	True Positive(TP)	False Positive(FP)
예측한 논스패머	False Negative(FN)	True Negative(TN)

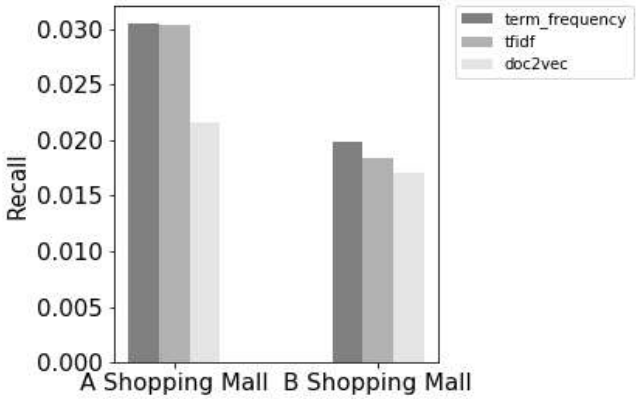
$$\text{Precision} = \frac{TP}{TP+FP} \qquad \text{Recall} = \frac{TP}{TP+FN}$$

2.3.5 실험 결과 및 평가

A 쇼핑몰에서 TF, TF-IDF, Doc2Vec은 각각 6,309명, 4,407명, 3,399명의 스패머를 탐지하였고, 세 알고리즘이 탐지한 스패머 수가 달라 이를 유사하게 맞춰주었다. 그 결과, TF는 3,327명 중 3,213명이 특징으로 뽑은 스패머와 일치하였고, TF-IDF는 3,305명 중 3,197명, Doc2Vec은 3,399명 중 2,277명이 일치하였다. 같은 방식으로 B 쇼핑몰에 대해 세 알고리즘으로 스패머를 탐지한 결과 각각 1,112명, 746명, 2,396명의 스패머를 탐지하였다. 마찬가지로, 세 알고리즘으로 탐지한 스패머의 수가 다르므로 스패머의 수를 조절하였다. TF는 785명 중 726명이 특징으로 뽑은 스패머와 일치하였고, TF-IDF는 746명 중 671명, Doc2Vec은 759명 중 621명이 일치하였다.



[그림 2] 알고리즘별 Precision 비교 그래프



[그림 3] 알고리즘별 Recall 비교 그래프

[그림 2]는 A 쇼핑몰과 B 쇼핑몰에 TF, TF-IDF, Doc2Vec 알고리즘을 적용하여 스패머를 탐지한 후 구한 Precision을 그래프로 나타낸 것이다. A 쇼핑몰의 Precision은 각각 0.9657, 0.9673, 0.6699였고, B 쇼핑몰에서는 각각 0.9248, 0.8995, 0.8182였다. [그림 3]은 각 쇼핑몰에서의 알고리즘별 Recall을 그래프로 나타낸 것이다. A 쇼핑몰에서 TF, TF-IDF, Doc2Vec의 Recall은 각각 0.0305, 0.0304, 0.0216였고, B 쇼핑몰에서는 각각 0.0199, 0.0184, 0.0170였다. Precision과 Recall을 통해 Doc2Vec보다 TF나 TF-IDF를 이용한 스팸 리뷰 탐지가 더 좋은 성능을 보임을 확인하였다. 이때, 스패머의 특징으로 탐지한 스패머들(ground truth)을 느슨한 조건으로 탐지하였으므로 Recall은 전체적으로 낮은 값을 기록하였다.

3. 결 론

본 논문에서는 Doc2Vec의 리뷰 내용 유사도 분석이 동일한 내용이 작성된 가짜 리뷰 탐지에는 어려움이 있다는 점을 검증하였다. TF, TF-IDF가 단어의 개수를 세어 벡터화하는 방식이라면 Doc2Vec은 단어의 벡터 거리를 기준으로 벡터화한다. 즉, TF와 TF-IDF는 단순히 토큰화된 리뷰들의 형태소의 종류와 개수만을 고려하는 반면, Doc2Vec은 리뷰 내용의 유사도를 고려하므로, 리뷰에 완전히 같은 내용이 없더라도 내용이 유사하면 스패머라고 판단한다.

스패머들의 특징으로 거른 스패머들을 ground truth로 설정하고 이들과 세 알고리즘으로 탐지한 스패머들을 비교하여 알고리즘의 성능을 분석하였다. 그 결과, Doc2Vec의 Precision과 Recall이 가장 낮았고, 이에 따라 리뷰의 내용을 바꾸는 노력을 들이지 않는 스패머들을 탐지하는 데는 한계가 있음을 확인하였다.

본 연구에서는 사용자별로 리뷰 내용의 유사도를 분석함으로써 Doc2Vec이 기계적으로 같은 리뷰를 작성하는 스패머들의 특성과는 맞지 않는다는 점을 확인하였다. 향후 연구로 여러 아이디를 이용해 동일한 내용을 작성하는 스패머를 탐지하는 방안을 모색해보고자 한다.

4. 사 사

본 연구는 과학기술정보통신부 및 정보통신기술진흥센터의 SW 중심대학지원사업 (2016-0-00021), ICT 연구과제 (2017-0-00396) 및 한국연구재단 이공분야 기초연구사업 (NRF-2016R1C1B2015135)의 지원을 받아 수행됨.

참고문헌

[1] SNEHAL DIXIT & A.J.AGRAWAL, "SURVEY ON REVIEW SPAM DETECTION", International Journal of Computer & Communication Technology ISSN, vol. 4, 2013

[2] Michael Crawford et al, "Survey of review spam detection using machine learning techniques", Jounral of BigData, 2015

[3] 민지수, 정혜리, 김가연, 유문상, 이경용, "국내 쇼핑 사이트의 스팸 리뷰 특징 탐지", KDBC, 2018 publish 예정, (in Korean)

[4] Quoc V. Le, Tomas Mikolov, "Distributed Representations of Sentences and Documents", 2014, https://arxiv.org/pdf/1405.4053.pdf

[5] http://konlpy.org/en/v0.4.4/api/konlpy.tag/

[6] http://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html

[7] http://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

[8] https://radimrehurek.com/gensim/models/doc2vec.html