의미 프레임 자질 기반 의견 스팸 분석

김성순, 장혁윤, 이성운, 강재우*
고려대학교 컴퓨터·전파통신공학과
e-mail : {seongkim, blackdragon1471, wanemoon, kangj}@korea.ac.kr

Deep Semantic Feature based Deceptive Opinion Spam Analysis

Seong-Soon Kim*, Hyeok-Yoon Jang**, Seong-Woon Lee** and Jaewoo Kang†
Dept. of Computer Science and Radio Communication Engineering, Korea University

요 약

소셜미디어의 급증과 함께 온라인 리뷰의 의존성이 급증하는 가운데 사용자의 올바른 의사결정을 저해하는 기만적 의견 스팸 이슈가 새롭게 주목받고 있다. 기존의 의견 스팸 연구는 실제 리뷰와 의견 스팸 간의 차이를 어휘, 품사 또는 감정단어와 같은 표면적 자질을 통해 설명하였으나 그들간의 의미적 연결관계는 고려하지 않았다. 본 논문에서는 1) 의미적 프레임 기반의 텍스트 분석 기법을 제안하고, 이를 바탕으로 2) 의견 스팸과 실제 리뷰간의 의미적 차이가 있음을 규명하며 3) 새로운 의미적 프레임 자질을 사용하여 기존의 의견 스팸 분류 성능을 향상시킬 수 있음을 보인다.

1. 서론

최근 소셜미디어의 발달과 맞물려 참여, 공유와 개방을 모토로 하는 Web 2.0 시대는 완연한 성숙기에 접어들었다. 웹 상에는 다양한 주제에 대한 수 많은 사용자들의 의견이 공유 및 전파되고 있으며 이러한 의견 정보는 실생활에서의 의사결정에 영향을 미치는 중요한 요소로 자리잡았다. [1]에 따르면 79%이상의 사용자들이 온라인상의 리뷰를 신뢰하며, 이를 실제 구매 행위에 반영한다는 사실이 조사되어 위 사실을 뒷받침한다.

한편, 온라인상의 타 사용자들의 의견(상품 리뷰)이 중요해지는 만큼 이를 비즈니스적으로 악용하는 사례가 점차 증가하고 있다. 예를 들어, 특정 점포의 사용자 평점을 높이기 위하여 해당 비즈니스의 이용 경험이 없는 제 3 자가 마치 실제로 만족스런 경험을 한 것처럼 꾸며낸 의견을 작성하는 경우가 다수 보고되었다. [2]

위와 같이 의도적으로 작성된 기만적 의견 스팸(Deceptive Opinion Spam, 이하 의견 스팸)은 사용자의 올바른 의사결정을 방해하며 웹상의 건전한 정보 유통을 저해하는 요소로 지적되고 있어 점점 그 심각성이 커지고 있는 반면 쉽게 해결되지 못하고 있다. 왜냐하면 대부분의 의견 스팸은 매우 교묘히 작성되어 심지어 사람 조차도 분간해 내기 어렵기 때문이다.[3]

[4]에서 의견 스팸 문제가 본격적으로 논의된 이후 최근 몇 년간 의견 스팸을 사람이 아닌 기계적 알고리즘을 통해 분류해 내려는 연구가 시도되었다. [3]는 n-gram 기반의 간단한 자질로 학습한 기계학습 기반 분류 모델로도 높은 분류 성능을 낼 수 있음을 보였다. 그러나 다른 연구와 마찬가지로 실제 리뷰와 의견 스팸간에 품사나 감정단어의 사용의 차이 등 얕은 깊이의 구문 분석(Shallow syntactic analysis) 외에 두 그룹이 의미적으로 다른 특성을 보이는 근본적인 이유에 대해서는 논의하지 않았다.

본 논문에서는 기존에 의견 스팸 연구에서 다루어지지 않았던 FrameNet(이하 프레임넷)[5] 기반의 의미론적 분석 기법을 제안한다. 의미적 프레임 단위 분석 기법을 통하여 (1) 기존의 개별 토큰 단위의 분석보다 한단계 더 나아가 문장 내에서의 단어간의 의미적 연결관계까지 파악할 수 있으며, (2) 의미적 프레임 발현 정도의 차이를 바탕으로 의견 스팸과 실제 리뷰 두 그룹간의 내용적 차이가 있음을 규명하였다. 또한 (3) 단어 토큰간의 의미적 연결관계를 기반으로 한 새로운 프레임 자질을 사용하여 기존의 의견 스팸 분류 성능을 향상시킬 수 있었다.

2. 관련 연구

2.1 FrameNet 개괄

FrameNet[9] 은 2003 년 Charles J. Fillmore 의 주도로 시작된 프로젝트로, '의미적 프레임 이론'(frame semantics theory)을 기반으로 하여 문장 내에 존재하는 사건/이벤트 및 해당 사건을 구성하는 객체(사람, 사물)간을 의미적 단위인 프레임으로 정형화한 일종의 사전이다. 예를 들어 다음의 문장 "I bought for a gift for her."의 경우 동사 'bought'는 '구입하다'라는 프레임 'COMMERCE_BUY' 를 촉발하는 중심 동사이며, 이와 문법적으로 연결된 주어 'I'와 목적어인 'gift'가

'COMMERCE_BUY'를 구성하는 핵심 의미 요소인 'Buyer'와 'Goods'에 대응된다. 위의 예제와 같이 프레임넷을 바탕으로 문장 내에 존재하는 의미 프레임을 추출하면 해당 문장이 어떠한 의미적 단위로 구성되어 있으며, 그것들의 연결 관계는 어떠한지에 대한 분석이 가능하다.

(그림 1) 'I bought a gift for her'문장의 프레임 분석 예

2.2 의견 스팸 관련 연구

[4]에 의해 본격적으로 의견 스팸에 대한 문제가 논의된 이후 현재까지 의견 스팸과 관련한 연구는 리뷰 단위 분석[3], 리뷰 작성자 단위 분석[6], 스패머 그룹 분석[7]과 같이 크게 3 가지 범주로 전개되어왔다. 본 연구에서는 사용자 메타정보를 고려하지 않고 리뷰 텍스트의 내용만을 분석 대상으로 하는 리뷰 단위 분석에 초점을 맞추었다.

리뷰 단위 의견 스팸 분석의 대표적 연구로는 [3]과 [8]을 들 수 있다. [3]는 Amazon Mechanical Turk(이하 AMT)를 통해 특정 호텔에 대한 경험이 없는 Tucker 가 해당 호텔의 직원이라 가정하고 이 호텔에 대한 긍정적인 평가를 남기도록 주문하여 의견 스팸 데이터를 수집하였다. 이렇게 만들어진 데이터셋을 사용하여 n-gram 이나 품사와 같이 비교적 간단한 자질 만으로도 사람의 성능을 능가하는 (정확도 85% 이상) 기계학습기반 분류 모델을 만들 수 있음을 실험을 통해 입증하였다. 이와 더불어 실제 리뷰에서는 명사, 형용사, 전치사, 한정사, 등위접속사 등이 우세하게 등장한 반면에 의견 스팸에서는 동사, 부사, 대명사 등의 어휘적 특성이 두드러지는 현상이 나타남을 보였다.

[8]은 [3]에서 배포한 AMT 데이터셋이 비즈니스 이용 경험이 없는 작성자가 작성한 의견 스팸의 형태만을 대표한다는 한계점을 지적하였다. 이를 극복하기 위하여 호텔, 레스토랑, 병원 도메인에 대하여 해당 비즈니스에 대한 전문적 지식과 경험이 있는 실제 직원을 섭외하여 긍정적 또는 부정적 평가를 작성하도록 한 '전문가 의견 스팸 데이터셋'을 추가하였다. 다시 말해, 경험의 유무와 관계없이 순수하게 실제 사용자와 의도적으로 작성된 리뷰 텍스트 상의 차이점만을 분석함으로써 의견 스팸을 대표할 수 있는 일반적인 특성을 찾고자 한 것이다. 실험 결과 품사 패턴과 LIWC[9] 의 '공간', '감성', '1 인칭 단수 대명사'가 기만적 의견 스팸에서 일반적인 특징적으로 발현되는 특성임을 발견하였다.

3. 제안하는 방법

이 장에서는 의미적 프레임 단위 분석을 위한 프레임 발현 정도의 비교하고, 각 그룹에서 나타난 대표적 프레임들에 대한 정성적 평가를 수행한다. 마지막으로 새로운 프레임 자질을 기계학습 모델의 자질로 사용하여 의미적 프레임의 유효성을 검증한다.

3.1 데이터셋 및 분석 도구

본 논문에서는 [3]의 Amazon Mechanical Turk(이하 MTurk)를 통해 작성된 데이터셋과 [8]의 전문가의 긍정 리뷰 데이터셋을 사용하였다. 두 데이터셋 간의 통일성을 위하여 연구 범위는 호텔 도메인으로 한정한다.

리뷰 문장의 n-gram 토큰화를 위해서는 Stanford Tokenizer[1]를 사용하였으며, 문장 분리기는 OpenNLP Sentence Detector[2]를, 자연어 문장에서 프레임넷 추출은 카네기멜론 대학의 SEMAPHORE 2.1[3] 버전을 사용하였다. 각 데이터셋과 SEMAPHORE 를 통해 추출된 프레임 통계는 아래 표와 같다.

	Deceptive		Truthful
	MTurk	Expert	
문서 갯수	400	140	400
총 프레임 갯수	17131	3641	19153
고유 프레임 갯수	467	322	463

<표 1> 데이터셋 및 프레임 추출 통계

3.2 분석 방법

이 절에서는 두 데이터셋의 프레임 발현 차이를 조사하기 위하여 '정규화된 프레임 발현빈도'(Normalized Frame Frequency, NFF)지표를 제안한다. NFF 는 다음과 같이 정의된다. 데이터셋 D, 프레임 f, 클래스 C={truth, deceptive}에 대하여 프레임 집합 $F_i = \{\forall f \; in \; D_i\}, i \in C$ 가 있을 때, 프레임 빈도 $f_q = frame\ occurrence\ in\ D_i$ where $f_q \in F$, $i \in C$ 와 같다. 특정 프레임 f_m에 대한 NFF 는

$$NFF_{D_j} f_m = f_q / \sum_{k=1}^{|F_j|} f_k, \; j \in C$$

와 같이 계산된다. 최종적으로 두 그룹에 등장한 각 프레임의 NFF 값의 차를 아래의 식으로 구한다.

$$\Delta NFF_{f_m} = NFF_{D_{deceptive}} f_m - NFF_{D_{truth}} f_m$$

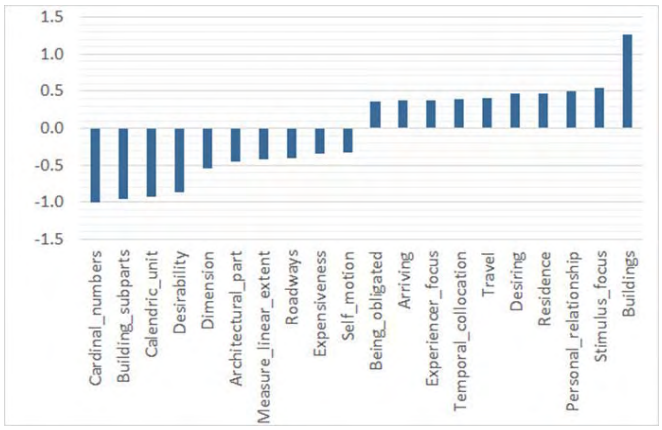
ΔNFF 도출 시 z-test 양측 검정을 통하여 두 모집단의 비율차이 검정을 수행한 결과, 두 그룹에서 NFF차이가 있음을 확인하였다 (p-value < 0.01).

[1] http://nlp.stanford.edu/software/tokenizer.shtml
[2] https://opennlp.apache.org
[3] http://www.ark.cs.cmu.edu/SEMAFOR/

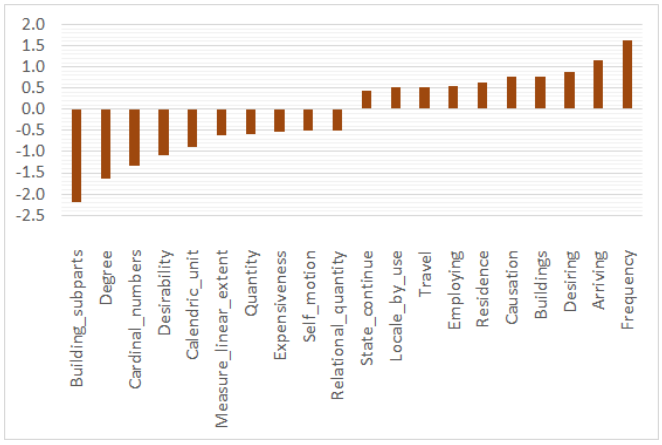
4. 실험

4.1 의견 스팸 및 실제 리뷰 그룹 간 ΔNFF 비교

[그림 2]와 [그림 3]은 각각 MTurk 와 실제 리뷰, 전문가 의견 스팸과 실제 리뷰 그룹 쌍의 각 프레임에 대하여 ΔNFF 값을 구한 뒤 이 값을 기준으로 정렬하여 양 극단의 Top10 개 프레임을 추려낸 결과이다. 만일 특정 프레임 f_m 에 대하여 ΔNFF_{f_m} 값이 음수이면 해당 프레임의 발현이 실제 리뷰보다 의견 스팸에서 더 많이 등장하였음을 뜻한다.



(그림 2) MTurk - Truthful 에서의 Top10 ΔNFF



(그림 3) Expert - Truthful 에서의 Top10 ΔNFF

(그림 2)와 (그림 3)의 좌측 그룹에서 보는 바와 같이 프레임 'Cardinal_numbers' 와 'Building_subparts' 는 실제 리뷰에서 더 빈번하게 등장하며, 'Buildings' 나 'Travel' 프레임은 실제 리뷰에 비해 의견 스팸에서 더 우세하게 등장한다는 사실을 알 수 있다. 이러한 특성은 기존의 연구들([3][7])에서 밝혀진 사실과 상당 부분 일치한다.

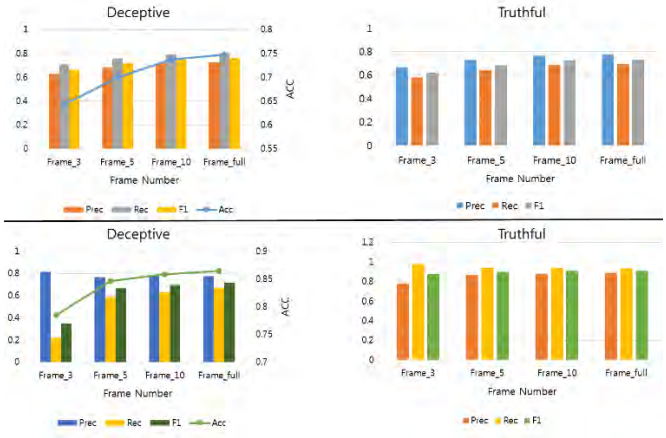
예를 들어, 의견 스팸은 작성자의의 개인적 경험이 결부되어 있기 때문에 장소에 대한 자세한 설명이 빈약한 경향이 있다. 이와 같은 이유로 [그림 2]와 [그림 3]에서 양의 값을 보이는 우측 프레임 그룹에서 '여행', '건물' ('Travel', 'Building') 등과 같이 서술 대상에 대한 피상적인 의미의 프레임들이 주로 포함되었음을 알 수 있다.

반면에 실제 리뷰는 작성자의 경험을 바탕으로 작성하였기 때문에 대상 비즈니스에 대한 자세하고 세부적인 내용들, 말하자면 '특정 날짜', '건물 내부', '가격이나 크기 또는 치수' (Cardinal_numbers', 'Dimension', 'Building_subparts', 'Calendric_unit', 'Expensiveness' 프레임)등의 내용을 의견 스팸에 비해 더 자주 언급하는 것으로 나타났다.

또 한가지 주목할 점은 의견 스팸 그룹에서는 읽는이로 하여금 해당 리뷰가 더욱 신뢰감을 주도록 하기 위하여 '배우자' 또는 '가족' 과 같이 개인적 관계((그림 2) 'Personal_relationship' 프레임)에 대한 언급을 자주 하는 경향이 있음이 관찰된다는 것이다. 이러한 현상은 [7]에서 언급한 것처럼 '1 인칭 단수 대명사' 를 자주 사용하여 개인의 경험에서 우러나와 작성된 내용임을 은연중에 드러내는 현상과 일맥상통한다.

4.2 Frame 자질을 사용한 기계학습 모델의 분류 성능 평가

본 절에서는 4.1 절에서 분석한 바와 같이 각 그룹에서 다르게 발현되는 프레임을 기계학습의 자질로 사용하여 의견 스팸과 실제 리뷰의 분류 실험을 전개한 내용을 서술한다.



(그림 4) 프레임 자질을 사용한 분류 모델 성능: 상단-AMT vs. Truthful; 하단-Expert vs. Truthful

(그림 4)는 프레임을 우리가 재현한 분류 모델의 단독 학습 자질로 사용한 결과이다. (그림 4) 상단에서 보는 바와 같이 Frame_3 세팅에서 ΔNFF 의 양 극단 값 3 개, 즉 단 6 개의 프레임 자질로도 무작위 선택 확률(0.63 > 0.5)을 상회하는 정확도를 얻을 수 있었다. 전문가 의견 스팸과 실제 리뷰 비교((그림 4) 하단) 실험에서는 프레임 자질 갯수를 10 에서 3 으로 줄임에도 정확도 값은 0.8 이하로 감소하지 않았다. 이로 미루어 볼 때 ΔNFF 로 얻어진 TopK 프레임 자질은 분류에 매우 효과적인 자질로 사용될 수 있음을 알 수 있다.

다음으로는 의미 프레임 자질이 기존에 연구된 의견 스팸 분류 모델의 성능을 향상시킬 수 있는지 알아보기 위하여 [3]의 방식을 재현하고 이를 베이스라인으로 설정하였다. 데이터셋은 5 겹 중첩 교차검증(5-fold nested cross validation) 방식으로 나눈 후 모델의 성능을 평가하였다. 분류기 구현 결과는 아래와 같다.

		Deceptive				Truthful		
	Features	Acc	Prec	Rec	F1	Prec	Rec	F1
Ott'11	Uni_svm	0.884	0.870	0.903	0.886	0.899	0.865	0.882
	Bi+_svm	0.896	0.891	0.903	0.897	0.901	0.890	0.896
Our Impl.	Uni_svm	0.870	0.868	0.873	0.870	0.872	0.868	0.870
	Bi+_svm	0.876	0.873	0.880	0.877	0.879	0.873	0.876

<표 2> [3]의 분류 모델 성능과 본 논문에서 재구현한 성능 비교

[3]에서 각 학습 집단 별 파라미터 및 세부 설정에 대한 정보가 주어지지 않아 본래 논문에서 밝힌 것과 같은 수치를 얻을 수는 없었으나, 재현한 분류 모델 성능과 근소한 차이를 보이고 있어 [3]의 방법을 동일하게 구현하였다고 간주한다.

		Deceptive				Truthful		
	SVM Features	Acc	Prec	Rec	F1	Prec	Rec	F1
Tucker vs. Truthful	BOW_full	0.870	**0.868**	0.873	0.870	0.872	**0.868**	0.870
	Frame5+BOW_full	**0.875**	0.859	**0.898**	**0.878**	**0.893**	0.853	0.872
	Frame5+BOW_250	**0.875**	0.864	0.890	0.877	0.887	0.860	**0.873**
Expert vs. Truthful	BOW_full	0.916	**0.857**	0.814	0.835	0.936	**0.953**	0.944
	Frame12+BOW_full	**0.920**	0.859	**0.829**	**0.844**	**0.941**	**0.953**	**0.947**

<표 3> 기존 BOW 자질에 프레임 자질 결합 후 분류 모델 성능 비교

<표 3>은 베이스라인 SVM 모델에 프레임 자질을 결합한 성능을 비교한 것이다. 두 데이터셋 모두에서 기존의 BOW(Bag-of-Word) 자질로 학습한 모델에 프레임 자질을 추가하였을 때 분류 정확도를 향상시킬 수 있었다. 특히 Tucker vs. Truthful 세팅에서 전체 unigram (약 5000 개)을 모두 사용하는 대신, Truthful 과 Tucker 각 셋에서 두드러지게 발현되는 unigram 을 ΔNFF 방식과 동일하게 추출하여 추가하였을 때 프레임 자질과 단 500 개 unigram 자질을 결합한 것 만으로도 BOW 기반 모델의 성능을 상회하는 결과를 얻을 수 있었다.

5. 결론 및 향후 연구

본 연구에서는 기존의 의견 스팸 연구에서 다루어지지 않았던 의미적 프레임 기반의 사용자 의견 문서 분석 방법을 제안하였다. 이를 통해 의견 스팸과 실제 리뷰 간의 차이를 이전 연구의 구문 레벨보다 한층 더 깊은 의미 단위에서 규명하였다. 또한 새로운 의미적 프레임 자질을 사용하여 기존의 의견 스팸 분류 모델의 성능을 향상시킬 수 있음을 보였다.

향후에는 호텔뿐만 아니라 레스토랑 및 다른 로컬 서비스 도메인 리뷰도 분석 대상으로 확장하여 본 제안 방법의 유효성을 검증하는 연구가 요구된다.

참고문헌

[1] "2013 Study: 79% Of Consumers Trust Online Reviews As Much As Personal Recommendations" (2013, June 26) Retrieved from http://searchengineland.com/2013-study-79-of-consumers-trust-online-reviews-as-much-as-personal-recommendations-164565

[2] "The Best Book Reviews Money Can Buy" (2012, Aug 25) Retrieved from http://www.nytimes.com/2012/08/26/business/book-reviewers-for-hire-meet-a-demand-for-online-raves.html?_r=0

[3] Ott, M., Choi, Y., Cardie, C., Hancock, J, T.: Finding deceptive opinion spam by any stretch of the imagination. In Proc. HLT'11. pp. 309-319 (2011)

[4] Jindal, N., Liu, B.: Opinion spam and analysis. In Proc. of WSDM. pp. 219-230 (2008)

[5] Lim, E., Nguyen, V., Jindal, N., Liu, B., Lauw, H, W.: Detecting product review spammers using rating behaviors. In Proc. CIKM'10. pp. 939-948 (2010)

[6] Mukherjee, A., Liu, B., Glance, N.: Spotting fake reviewer groups in consumer reviews. In proc. WWW'12. pp. 191-200 (2012)

[7] Li, J., Ott, M., Cardie, C., Hovy, E.: Towards a General Rule for Identifying Deceptive Opinion Spam. In Proc. ACL'14. pp. 1566-1576 (2014)

[8] Tausczik, Y. R., Pennebaker, J, W.: The psychological meaning of words: LIWC and computerized text analysis methods. Journal of Language and Social Psychology. 29 (1), pp. 24-54 (2010)

[9] Baker, C. F., Fillmore, C. J., Cronin, B.: The Structure of the Framenet Database. International Journal of Lexicography. 16 (3), pp. 281-296 (2003)