SNS(트위터)를 활용한 재난 및 위기상황 인식에 관한 연구

최연호, 현상엽, 신연순
동국대학교 컴퓨터공학과
dusgh4321@naver.com, reserved-one@naver.com, ysshin@dongguk.edu

A Study on the Perception of Disaster and Crisis Using SNS(Twitter)

YeonHo Choi , SangYeop Hyun, Younsoon Shin
Dept. of Computer Science and Engineering, Dongguk University

요 약

재난 및 위기상황이 발생하면 해당 상황을 신속하고 정확하게 파악해야 많은 사람들을 구조할 수 있다. 본 논문은 SNS에서 재난 및 위기 상황을 정확하게 인식하는 연구를 진행한다. 텍스트 정규화, 워드 토큰화, 단어 임베딩 과정을 통해 전처리를 진행하고 키워드와 여러 특징들을 뽑아 SVM classifier를 사용하여 분류 작업을 실시한다. 실험결과 재난과 연관이 있는 경우에 해시태그의 빈도수, URL 빈도수, 두 키워드간의 거리가 다른 특징들의 조합보다 더 좋은 결과를 나타내었다.

1. 서론

재난 및 위기상황이 발생하면 해당 상황을 신속하고 정확하게 파악해야 많은 사람들을 구조할 수 있다. 재난 및 위기상황에 대처하기 위해서는 그 상황에 대해 분석하기 위한 데이터와 알고리즘이 필요하다. 이를 토대로 정확한 위기상황을 파악하고 경보를 주어 피해자들을 대피시킬 수 있다. 이와 관련된 많은 연구가 진행되었고 그중에 SNS를 기반으로 한 연구들이 있다.

본 연구는 SNS에서 재난 및 위기상황, 특히 지진에 대한 정보를 신속하고 정확하게 파악할 수 있는 방법에 관한 것이다. 기존 [1]의 연구에서는 하나의 키워드만 추출해 재난 상황을 판단하였으며 그 결과 Recall은 좋게 나왔지만 Precision과 F1-score가 좋지 않게 나왔다. 본 연구에서는 연관이 있는 키워드를 하나 더 추출하고 feature를 추가하여 Precision과 F1 score를 높이는 방안을 연구한다. 우선 SNS 중에 twitter data를 사용하여 Text normalization, Word tokenization, Stop-word removal 3가지 전처리 과정을 진행한다. 이후 feature extraction으로 총 7가지 특징을 추출한다. 이 과정에서 연관 키워드는 단어 임베딩을 사용하여 찾아낸다. 마지막에 Precision, Recall, F1-score를 구하여 결과에 대해 판단한다.

2. 관련 연구

이 절에서는 연관 단어를 찾아내는 단어 임베딩, Support Vector Machine Classifier를 소개한다.

2.1 워드 임베딩 (Word embedding)

단어를 밀집벡터의 형태로 표현하는 방법을 워드 임베딩(word embedding) 또는 임베딩 벡터(embedding vector)라고 한다. one hot vector와 달리 저차원이며 1과 0 외에도 실수 값을 사용하고 data로부터 학습된다. 대표적인 방법으로 Word2vec이 있다. 단어의 의미를 여러 차원으로 분산하여 표현하고 비슷한 위치에서 등장하는 단어들은 비슷한 의미를 가진다. 즉, 단어 간 유사도를 계산할 수 있다. 본 논문에서는 Word2vec을 사용하여 키워드와 연관된 다른 키워드를 찾는다.

2.2 Support vector machine

SVM은 classification과 regression에 응용할 수 있는 supervised learning이다. support vector와 hyperplane을 이용해서 주로 classification을 수행한다. 어느 한 쪽에 치우치지 않게, 양쪽 데이터와 균등한 위치에 분류 기준을 세워야 한다. Margin을 최대로 함으로써 새로운 data에 대한 분류를 정확히 할 수 있다[2]. 본 논문에서는 SVM classifier를 사

용하여 재난 및 위기상황 여부를 학습하고 테스트하여 결과를 도출한다.

3. 구조

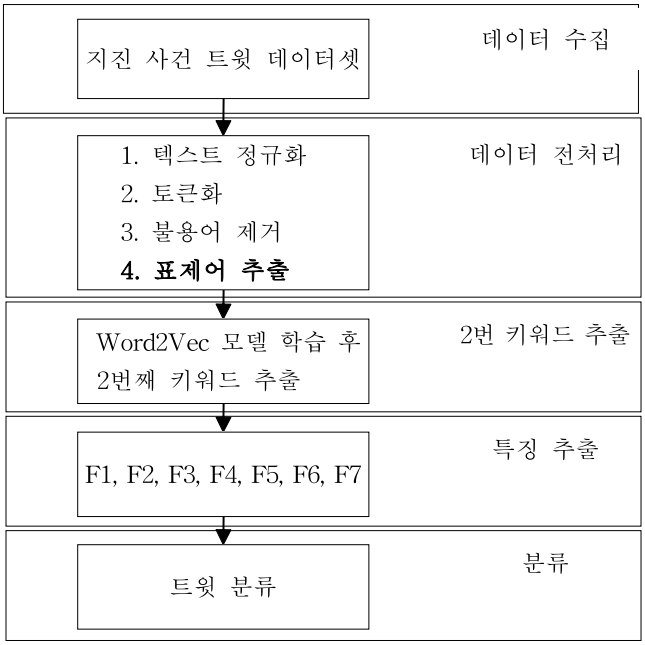
기존 솔루션에서 제시하는 텍스트 전처리 과정과 실험에 사용될 특징의 추출 과정을 보완하여 제안하는 알고리즘을 통해 트윗 데이터는 텍스트 정규화 및 토큰화, 불용어 제거와 표제어 추출을 포함하여 전처리된다. 기존 5개의 특징에 더해 추가된 키워드의 위치와 기존 키워드와 추가된 키워드 간의 거리 특징은 다른 특징들과 동일하게 전처리된 데이터에서 추출된다. 이러한 특징들은 모두 트윗이 특정 재난 사건과 관련이 있는지 없는지 탐지하기 위해 SVM 분류기에 제공된다. 제안된 지진 사건 탐지 방법은 아래의 알고리즘1과 같다.

알고리즘1 제안된 재해 사건 탐지 알고리즘

(1) 지진 사건 트윗 데이터셋 수집
(2) X = {tweet_1, tweet_2, ... , tweet_n}.
(3) 집합 X의 각 트윗 x에 대해, F1, F2, F3, F4, F5, F6, F7 특징을 추출.
F1 - 트윗 내 1번째 키워드의 위치
F2 - 트윗의 길이
F3 - 트윗 내 해시태그(#)의 빈도수
F4 - 트윗 내 사용자 언급(@)의 빈도수
F5 - 트윗 내 URL 빈도수
F6 - 트윗 내 2번째 키워드의 위치
F7 - 1, 2번 키워드 간의 거리
(4) SVM을 사용하여 지진 사건과 관련되거나 관련 없는 트윗을 분류.
(5) 다른 특징 조합으로 4번 과정을 반복.

제안된 시스템의 구조는 그림 1과 같다. 기존과 비교해 제안된 시스템의 단계는 다음과 같다:
(1) 데이터 수집
(2) 데이터 전처리
(a) 텍스트 정규화 및 토큰화
(b) 불용어 제거
(c)* 표제어 추출
(3)* 특징 추출(7가지),
(4) 분류

통계적 특징은 트윗 데이터 구성에 대한 구조와 키워드를 통한 정확도를 높이기 위하여 활용되며 일부 특징은 새로운 데이터 분류에 대한 쏠림 현상을 방지하기 위한 장치로 활용됨을 명시한다.



학습 및 실험에 사용되는 트윗 데이터는 지진과 관련된 사건을 나타내는 상황, 비상황적 정보와 지진과 무관한 사건을 나타내는 비상황적 정보가 모두 포함되어 있다. 수집된 데이터는 각각 텍스트 정규화, 토큰화, 불용어 제거, 표제어 추출을 통해 전처리된다.

텍스트 정규화 프로세스는 수집된 모든 트윗을 소문자로 변환하는데, 이는 Earthquake, EarthQuake, EARTHquake 등 같은 단어가 다른 형식으로 나타날 수 있기 때문이다. 앞서 언급된 식별된 키워드에 대해 텍스트 정규화가 적용되지 않으면 Earthquake, EarthQuake, EARTHquake 등의 단어는 별도의 단어로 간주한다. 따라서 텍스트 정규화는 단어의 위치를 식별하는 데에 필수적이다.

정규화 프로세스 이후에는 토큰화가 적용되는데, 이는 트윗을 여러 단어로 나누는 과정이다. (예시) Earthquake occurs in the Nepal.에서 “Earthquake”, “occurs”.“in”,“the”,“Nepal”은 예제 트윗의 토큰이다. 불용어는 트위터의 최대 트윗에서 사용할 수 있는 일반적인 단어이다. 이러한 단어는 대상 사건을 탐지하는 동안 트윗에서 어떠한 의미를 제공하지 않기에 데이터로부터 제거해도 무방하다. 예시 트윗의 불용어는 “in”,“the”등이 있으며 그 외에도 다른 트윗에서 발견할 수 있는 불용어의 예시로는 다음의 “More”,“so”,“that”,“is”,“a”,“by”,“of” 등이 있다. 이러한 단어들은 모두 제거된다.

트윗 번호	트윗	지진 사건 연관성
1	For those of you unware... An earthquake in Nepal just killed 4,000 pelple.	연관
2	#Earthquake may have an effect on #hydroelectric facilities [URL] #AltaaqaGlobal #Renewable #Rental #Temporary #Genset	연관
3	So awful hearing about the devastation and amount of people killed due to the earthquake in Nepal [URL]	연관 없음
4	Mirror image video of Mexico #earthquake is shared as footage from Nepal swimming pool.	연관 없음

표 1 지진 사건과 연관 유무에 따른 트윗의 예시

트윗 번호	트윗 내 1번째 키워드 위치	트윗 길이	트윗 내 해시태그 (#)의 빈도수	트윗 내 사용자 언급 (@)의 빈도수	트윗 내 URL 빈도수	트윗 내 2번째 키워드 (killed) 위치	1,2번 키워드 간의 거리
1	2	6	0	0	0	4	2
2	1	11	7	0	1	0	1
3	8	10	0	0	1	6	2
4	5	10	0	0	0	0	5

표 2 예시 트윗에서 추출한 7가지 특징

전처리 과정을 거친 후의 결과는 그림 2와 같이 나타난다.

제안 시스템의 세 번째 단계는 특징 추출이다. 특징들은 지진 사건을 탐지하는 데에 중요한 역할을 한다. 본 연구에서는 7가지의 통계적 특징을 추출하여 다음과 같이 제시한다.

(1) 트윗 내 1번째 키워드의 위치(F1). 이는 전처리된 트윗 내에서 1번째 키워드 발생을 나타낸다.

(2) 트윗의 길이(F2). 불용어를 제거한 후의 트윗 단어 수를 카운트한다.

(3) 트윗 내 해시태그(#)의 빈도수(F3). 트윗에서 해시태그가 발생하는 횟수를 카운트한다. 해시태그는 단어 앞에 등장하는 레이블이며 트윗의 주제를 식별하는 데에 사용된다.

(4) 트윗 내 사용자 언급(@)의 빈도수(F4). 트윗에서 사용자가 언급된 횟수를 카운트한다. @로 표시하며 이는 다른 사용자에게 회신하는 데에 사용되는 사용자 이름보다 우선 한다.

(5) 트윗 내 URL 빈도수(F5). 트윗에 URL이 표시되는 횟수를 카운트한다. URL은 이미지, 비디오 등의 형태로 추가 정보를 제공한다.

(6) 트윗 내 2번째 키워드의 위치(F6). 이는 전처리된 데이터로 학습된 Word2Vec 모델에 1번째 키워드를 통해 추출한 키워드 중 가장 유사도가 높다고 판단 되어진 2번째 키워드이며, 트윗 내에서 2번째 키워드가 발생했음을 나타낸다.

(7) 1, 2번 키워드 간의 거리(F7). 트윗 내에서 발생한 1,2번째 키워드 간의 거리를 계산한 것이다. 단, 키워드가 하나만 존재한다면 해당 키워드의 위치(F1,F6)와 동일한 값을 가진다.

1. [‘rt’,‘@ochaasiapac:’,‘#nepalquake’,‘estimated’,‘4.6’,‘million’,‘people’,‘exposed’,‘#earthquake’,‘shaking’,‘[url]’]
2. [‘nepal’,‘earthquake:’.‘devastation’,‘map’,‘images:’,‘satellite’,‘image’,‘map’,‘[url],’#news’]

그림 2. 전처리 후 결과의 예시

7가지 특징은 표 1에 표시된 예제 트윗에서 추출한 것으로, 추출한 특징별 분류는 표 2에 나와 있다. 특징 F1, F6에 해당하는 키워드는 1번째는 ‘earthquake’ 2번째는 ‘killed’를 예시로 하였다. 모든 통계적 특징을 추출한 후, 그것들은 지진 사건과 관련이 있거나 없는 트윗을 분류하기 위해 SVM 분류기에 주어진다.

4. 실험 및 분석

이 절에서는 제안된 알고리즘의 실험 및 분석에 사용되는 데이터 셋과 키워드 선택, 그리고 그에 대한 성능 분석에 관하여 설명한다.

4.1 데이터 셋

제안된 알고리즘은 파이썬 언어로 구현되었으며 2015 네팔 대지진 데이터셋으로 수집된 트윗으로 실험을 진행하였다. 훈련에 사용된 트윗은 지진 키워드를 포함, 불포함 총 6,899개의 트윗이며, 이 중 3,293개는 지진과 관련이 있고 나머지 트윗은 지진과 관련이 없다. 테스트에 사용된 트윗은 지진 키워

드를 포함, 불포함 총 3,479개의 트윗이며, 이 중 1,636개는 지진과 관련이 있고 나머지 트윗은 지진과 관련이 없다. 트윗 데이터에는 경고, 조언, 부상, 실종 등 다양한 범주의 트윗이 포함되어 있으며 이러한 범주들의 트윗이 결합되고 지진 사건과 관련된 것으로 간주되는 데이터를 포함한다.

4.2 성능 평가

F1, F2, F3, F4, F5, F6, F7 특징들의 조합은 분류를 위해 SVM에 주어지며, 제안된 시스템의 성능은 정밀도(precision), 재현율(recall), f1-score의 세 가지 매개 변수와 정확도(accuracy)로 평가할 수 있다. 분류기에 사용된 매개변수 및 2번째 키워드 추출을 위해 자연어 처리에 사용된 Word2Vec 모델의 매개변수 또한 표 2에 나와 있다. 특정한 특징의 사용에 따라 높은 사건 탐지율 혹은 관련 없는 지진 사건의 탐지로 인한 낮은 정밀도 등, 재현율과 정밀도 사이의 조화 평균을 나타내는 f1-score와 전체 데이터에 대하여 분류의 정확성을 나타내는 정확도를 본 연구에서는 결과 분석의 중점적 요소로 사용한다.

4.3 실험

본 연구에서는 네팔 대지진 데이터셋 트윗을 사용하고, 1번째 키워드인 'earthquake'외에 2번째 키워드를 Word2Vec 모델을 통해 연관성이 있는 키워드를 추가로 설정하여 진행하였다. 2번째 키워드는 'aftershock','friend'로 설정하여 실험을 진행하였으며 이는 단어별로 학습 및 테스트에 사용하기 위한 트윗의 수는 최소 300개 이상이어야 한다는 가정을 바탕으로 선정되었다. 트윗 표본의 수가 키워드 조합별 차이가 있기 때문에 정밀도, 재현율, f1-score가 조합에 따라 점수 차이를 보이지만 본 연구는 7가지의 특징 조합 중 가장 성능이 우수한 경우에 대한 분석을 목표로 하기에, 1번째 키워드는 'earthquake'로 동일하며 2번째 키워드의 변화에 대해 가장 높은 성능을 보이는 특징 조합을 분석 결과로서 그 중, f1-score와 정확도가 높은 조합을 우선 시 한다.

표 3, 4는 7가지 특징 중 f1-score와 정확도가 높은 여러 조합을 2번째 키워드 'aftershock'와 'friend'에 따라 성능을 비교한 결과값이 제시되어 있다. 상대적으로 지진 사건과 연관이 높은 'aftershock'와 낮은 'friend', 두 키워드를 통해 결과값에 대한 차이를 확인할 수 있다. 본 연구에서는 표본의 차와 별개로 추출한 특징의 조합에 따른 f1-score와 정확도의 향상을 주된 목적으로 한다.

특징	재현율	정밀도	f1-score	정확도
F1,F3	100	77	**87**	**77**
F1,F3,F5	100	77	**87**	**77**
F1,F4,F5	93	77	84	74
F3,F5,F6,F7	93	78	85	75

표 3 2번째 키워드 'aftershock'와의 특징 조합 경우(%)

특징	재현율	정밀도	f1-score	정확도
F1,F6	44	51	47	59
F4,F6	51	59	55	65
F4,F5,F6	54	61	**57**	**67**

표 4 2번째 키워드 'friend'와의 특징 조합 (%)

5. 결론

본 논문에서는 다양한 통계적 기능 조합과 SVM 분류기를 사용, 추가 키워드 및 특징 분류를 통해 기존 연구 대비[1] 향상된 재난과 관련된 트윗을 탐지하는 방법을 제안한다. 지진과 관련된 하나의 핵심 키워드만을 사용하는 것이 아닌 그에 따른 연관성 높은 키워드를 추가로 탐색하여 추출하고, 2번째 키워드로 활용하여 그 위치와 거리를 새로운 특징 조합으로 사용하였으며 이를 통한 실험 결과에서, 분석을 통해 상대적으로 지진 사건과 연관이 높아 표본이 높은 2번째 키워드의 사용 경우는 해시태그의 빈도수와 URL의 빈도, 1,2번째 키워드간 거리가 다른 조합보다 더 좋은 결과를 나타냄을 알 수 있고, 상대적으로 지진 사건과 연관이 낮아 표본이 적은 2번째 키워드의 사용 경우는 사용자 언급 빈도수와 2번째 키워드의 위치와의 조합이 더 좋은 결과를 제공함을 알 수 있다. 본 연구서 제안된 추가적 특징과의 조합이 특정 상황에 대해 기존 연구 대비 더 좋은 성능을 보임을 확인할 수 있으며 향후, 추가 특징 추출의 확장과 호환을 통하여 특정 재난 사건 외 다른 재난 사건에도 적용할 수 있을 것이다.

ACKNOWLEDGMENT

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학지원사업의 연구결과로 수행되었음.
(2016-0-00017)

참고문헌

[1] Madichetty Sreenivasulu and M. Sridevi, "Comparative Study of Statistical Features to Detect the Target Event During Disaster", BIG DATA MINING AND ANALYTICS, pp 121 - 130, June 2020.

[2] H. T. Sueno, B. D. Gerardo, and R. P. Medina, "Multi-class document classification using support vector machine (SVM) based on improved naïve bayes vectorization technique," Int. J. Adv. Trends Comput. Sci. Eng., vol. 9, no. 3, p3939, 2020.