Journal of Korea Multimedia Society Vol. 23, No. 9, September 2020(pp. 1181-1190)
https://doi.org/10.9717/kmms.2020.23.9.1181

텍스트 스트리밍 데이터에서 텍스트 임베딩과 이상 패턴 탐지를 이용한 신규 주제 발생 탐지

최세목[†], 박정희[††]

Emerging Topic Detection Using Text Embedding and Anomaly Pattern Detection in Text Streaming Data

Semok. Choi[†], Cheong Hee Park[††]

ABSTRACT

Detection of an anomaly pattern deviating normal data distribution in streaming data is an important technique in many application areas. In this paper, a method for detection of an newly emerging pattern in text streaming data which is an ordered sequence of texts is proposed based on text embedding and anomaly pattern detection. Using text embedding methods such as BOW(Bag Of Words), Word2Vec, and BERT, the detection performance of the proposed method is compared. Experimental results show that anomaly pattern detection using BERT embedding gave an average F1 value of 0.85 and the F1 value of 1 in three cases among five test cases.

Key words: Anomaly Pattern Detection, Text Streaming Data, Text Embedding, Word Embedding

1. 서 론

스마트폰과 같은 모바일 기기의 발전과 온라인 매체의 발달로 인해 실시간으로 발생하는 스트리밍 데이터의 양이 크게 증가하고 있고 스트리밍 데이터 마이닝 기법들이 연구되고 있다. 특히, 정상적인 데이터 패턴을 벗어나는 이상치 탐지와 이상 패턴 발생 탐지에 대한 연구는 네트워크 침입 탐지, 불법 카드 사용자 예측, 시스템 이상 발생 탐지 등 여러 분야에 적용되는 기술이다[1]. 이상치 탐지는 개별적인 데이터 샘플이 이상데이터인지 아닌지 예측하는 데 반해서, 이상 패턴 탐지는 데이터 스트림에서 정상을 벗어나는 데이터 생성 패턴 발생을 탐지하는 것을 목표로 한다[2]. 최근에는 정상 학습 데이터를 이용한 이상치 탐지 모델 구성과 이상치 예측에 의한 이진값 스트리밍 데이터로의 변환에 기반한 이상 패턴 탐지 방법이 제안되었고, 스마트 미터기로부터 측정되는 전력사용량 스트리밍 데이터에서 전기 도둑 탐지에 적용되었다[3,4].

정형 또는 비정형 등의 다양한 유형의 데이터 생성과 저장이 가능해짐에 따라 이에 맞는 데이터 분석 방법의 개발도 꾸준히 이루어지고 있다. 특히, 트위터, 페이스북과 같은 소셜 네트워크 서비스와 네이버, 다음과 같은 포털 사이트에서 제공하는 카페와 같은 각종 커뮤니티 사이트에서는 대량의 텍스트 데이터가 발생한다. 이러한 데이터들을 분석하고 처리하기 위해서 텍스트 마이닝의 중요성이 부각되고 있고, 관련 연구가 활발히 이루어지고 있다. 텍스트 스

※ Corresponding Author : Cheong Hee Park, Address: (34134) Daehak-ro 77, Yuseong-gu, Daejeon, Korea, TEL : +82-42-821-6293, FAX : +82-42-822-4997, E-mail : cheonghee@cnu.ac.kr
Receipt date : Aug. 10, 2020, Revision date : Sep. 3, 2020
Approval date : Sep. 7, 2020

[†] Division of Computer Convergence, Chungnam National University (E-mail : semok95@daum.net)
[††] Division of Computer Convergence, Chungnam National University
※ This research was supported by Chungnam National University.

트리밍 데이터에서 이상 패턴 탐지와 관련된 연구로는 First Story Detection(FSD)가 있다. FSD는 스트리밍 텍스트 데이터에서 새로 발생하는 주제에 관한 첫 번째 문서를 찾는 것을 목표로 한다[5]. Topic Detection and Tracking(TDT)은 브로드캐스트 뉴스(broadcast news)를 각각의 스토리로 분할하기, 클러스터 내에 연관된 주제의 스토리들이 있도록 주제 클러스터들을 구성하기, 몇 개의 스토리로 구성된 주제가 주어졌을 때 스트리밍 데이터에서 같은 주제에 속하는 스토리들을 찾아내기 등의 세 가지 요소로 구성된다[6]. 이에 반해 이상 패턴 탐지는 정상적인 데이터 분포에서 발생한다고 간주되는 데이터 스트림에서 정상을 벗어나는 데이터들이 급작스럽게 많이 발생하게 되는 시점을 탐지하는 것이다. 예를 들어 스트리밍 텍스트 데이터에서 이전에 나타나지 않았던 새로운 주제에 관한 텍스트들의 발생을 탐지하는 것을 의미한다.

비정형 데이터인 텍스트를 처리하기 위해서는 컴퓨터가 인식할 수 있는 특징 벡터로 변환해야 할 필요가 있다. 텍스트를 벡터 형태로 표현하는 전통적인 방법으로는 Bag of Words(BOW)가 있다. 최근에는 단어 사이의 연관 관계를 고려한 단어 임베딩 모델인 Word2Vec[7]과 언어 처리 신경망 모델인 BERT (Bidirectional Encoder Representations from Transformers)[8] 등 임베딩 방법들이 활발히 연구되고 있다.

본 논문에서는 텍스트 임베딩 방법을 사용하여 텍스트 스트림에 있는 텍스트에 대한 벡터 표현을 계산하여 벡터 데이터 스트림으로 변환하고, 이상 패턴 탐지 방법을 적용하여 데이터 스트림에서 새로운 주제 발생을 탐지하는 방법을 제안한다[9]. BOW, Word2Vec, BERT와 같은 다양한 텍스트 임베딩 방법들을 사용하여 텍스트 데이터 스트림에서 이상 패턴 탐지 성능을 비교한다. 본 논문의 구성은 다음과 같다. 2절에서는 텍스트 데이터에서 신규 주제 발생 탐지와 관련된 연구를 살펴보며 3절에서는 대표적인 텍스트 임베딩 방법들에 대해서 설명한다. 4절에서는 텍스트 임베딩과 이상 패턴 탐지 방법을 활용한 신규 주제 탐지 방법에 대해 소개한다. 5절에서는 실험 데이터와 실험 결과를 다루고, 6절에서는 본 논문의 결론을 맺고 마무리한다.

2. 관련연구

텍스트 데이터 스트림에서 새로운 주제에 관한 첫 번째 문서(스토리)를 탐지하는 First Story Detection(FSD)는 본 논문에서 제안한 신규 주제 발생 탐지 방법과 가장 밀접하게 연관된 연구주제이다. FSD의 기본적인 접근 방법은 새로운 문서의 가장 가까운 이웃을 그 이전의 택스트 스트림에서 구하고, 새로운 문서와 가장 가까운 문서와의 거리를 이용해 새로운 스토리인지 결정하는 것이다. 가까운 이웃을 계산하는 복잡도를 낮추기 위해 LSH를 사용하거나, 의미상 연관된 요소들로 문서를 확장하는 등의 방법으로 성능을 향상시켰다[5,10,11,12]. 논문 [5]에서는 텍스트 데이터를 Tf-Idf(Term frequency-Inverse document frequency)를 사용하여 임베딩 벡터로 변환하고, 새로운 텍스트 데이터가 주어졌을 때, 1-NN 탐색을 통해 새로운 텍스트와 가장 가까운 데이터를 찾아 거리를 계산한 뒤, 거리가 사전에 설정된 임계값을 초과하는 경우 해당 텍스트를 신규 주제로 분류하는 모델을 제안하였다. 논문 [10]은 논문 [5]의 시스템에 Locality Sensitive Hashing(LSH)를 적용하여 유사한 텍스트들을 군집화하고, 새로운 텍스트가 발생했을 때, 앞서 들어온 모든 텍스트와 비교하는 대신 유사한 군집에 포함된 텍스트들과만 비교하는 방법으로 최적화한 시스템을 제시하고, 이를 실제 대용량 twitter 데이터 처리에 적용하였다. 논문 [11]은 같은 의미를 가지는 서로 다른 단어들을 사용하는 경우에 대한 처리를 위해 단어 임베딩을 사용하여 단어 쌍을 생성하여 단어 불일치로 인한 분류 문제를 해결하였다. 논문 [12]에서는 각 문서에서 개체-관계를 추출함으로써 문서를 표현하고, 이를 사용하여 유사도 계산을 수행하였다. 논문 [13]은 SNS에서 발생하는 짧은 길이의 텍스트 데이터로부터 특징 정보를 추출하고 마르코프 논리 네트워크를 사용하여 텍스트 데이터의 카테고리 분류를 수행하였다. 그러나, FSD는 새로운 주제의 첫 번째 문서를 찾는데 목표를 두는 반면에 본 논문의 신규 주제 탐지 방법은 새로운 주제의 문서들의 발생 빈도가 갑작스럽게 증가하는 시점을 탐지하고자 하므로, 두 연구는 목적과 성능 평가 방법에서 차이가 있다.

새로 발생하는 주제를 탐지하는 연구는 트윗 스트림에서 화제가 되는 사건의 발생을 탐지하는 데 활발히 적용되었다. 트위터에서 주제를 탐지하기 위해 제

안된 방법들 중에서, 단어 중심(feature-pivot) 토픽 탐지 방법은 동시에 등장하는 단어 그룹을 찾아내어 주제를 탐지하는 방법이며[14,15], 문서 중심(document-pivot) 탐지 방법은 문서들 간의 유사도를 비교하고, 문서들을 군집화함으로써 탐지하는 방법이다[10,16]. 확률 토픽 모델(probabilistic topic model)은 주제에 대한 단어의 분포와 문서에 대한 주제의 확률 분포를 추론하여 주제를 탐지한다[17]. 논문 [14]에서는 트위터 스트림을 처리하면서 짧은 시간에 등장 빈도가 급상승한 단어들을 추출하고, 이를 그룹화하여 스트림 내의 신규 주제를 탐지하였다. 논문 [15]에서는 트위터 스트림을 시간 단위로 나눈 뒤, High Utility Pattern Mining(HUPM)을 사용하여 현재 타임 윈도우에서 바로 직전 타임 윈도우에 비해 높은 빈도와 높은 유틸리티를 가지는 키워드를 추출하는 방법으로 새로 발생한 주제를 탐지하였다. 논문 [18]에서는 GPU를 사용한 병렬 연산을 통해 Tf-Idf를 계산하고, 이를 사용하여 텍스트를 <term, document> 쌍으로 변환한 다음, 상위 k개의 고빈도 쌍을 추출하는 방법으로 주제를 탐지하고 분석하였다. 하지만 기존 방법들은 매개 변수 설정이 어렵고 단어 집합이나 문서 집합으로 표현되는 주제를 탐지하는데 반해, 본 논문에서 제안하는 방법은 새로운 주제 발생 시점을 탐지한다는 점에서 차이가 있다.

3. 텍스트 임베딩 방법

3.1 BOW(Bag Of Words)

전통적인 텍스트 임베딩 방법으로는 Bag of Words (BOW)가 있다. BOW는 텍스트 데이터에서 등장하는 각 단어의 등장 빈도를 기록하여 표현하는 임베딩 방법이다. 단어의 등장 빈도만을 사용하기 때문에 텍스트에서 등장 빈도가 적지만 중요한 의미를 가지는 단어가 무시될 수 있는 한계가 있다. 이러한 한계를 극복하기 위한 방법으로 단어가 등장하는 문서의 수를 이용하여 가중치를 주는 Tf-Idf를 적용한다. 문서 d에서 단어 t의 등장 횟수를 나타내는 단어 빈도 $tf(t,d)$, 전체 문서 집합 D에서 단어 t가 등장하는 문서의 수인 문서 빈도 $df(t,D)$와 이것의 역을 취한 역문서빈도 $idf(t,D)=log(|D|/df(t,D))$를 사용하여 문서 내의 각 단어에 식(1)과 같이 가중치를 부여한다[19].

$$Tf-Idf(t,d,D)=tf(t,d)\times idf(t,D) \quad (1)$$

Tf-Idf는 단어가 특정 문서에서만 자주 등장할 경우 해당 단어는 높은 가중치를 가지며, 모든 문서 전체에 균일하게 등장할수록 가중치는 매우 작게 된다. Tf-Idf에 의한 문서 임베딩은 문서의 길이에서의 편차를 줄이기 위해 L_2-norm 길이를 1로 하는 정규화 과정을 수행한다.

3.2 Word2Vec

Word2Vec은 Mikolov 등이 제안한 단어 임베딩 방법으로, 입력 계층, 출력 계층과 하나의 투영(projection) 계층으로 구성되는 신경망 모델이다[7]. Word2Vec 모델 구조 중 하나인 skip-gram 모델은 목표 단어를 입력으로 하여 목표 단어 주변의 단어들을 예측하는 확률을 최대화하는 방향으로 학습을 수행한다. Fig. 1은 skip-gram 모델의 구조를 보인다.

투영 계층에서 단어 사전에 있는 단어들의 투영 행렬 W를 학습하게 되고 이것을 이용해서 단어들의 임베딩 벡터를 구한다. 그러나 Word2Vec 모델은 학습 과정에서 존재하지 않았던 새로운 단어가 발생하는 경우 해당 단어의 벡터 표현을 구할 수 없는 Out-of-vocabulary(OOV)의 문제가 있다. 이러한 한계를 극복하기 위해 제안된 모델로 fastText[20] 모델이 있다. fastText 모델에서는 단어를 bag of character n-gram으로 표현하고, 이를 사용하여 단어 벡터의 학습을 진행한다. 단어의 벡터 표현을 해당 단어에 포함된 n-gram 벡터들의 합으로 표현함으로써 fast

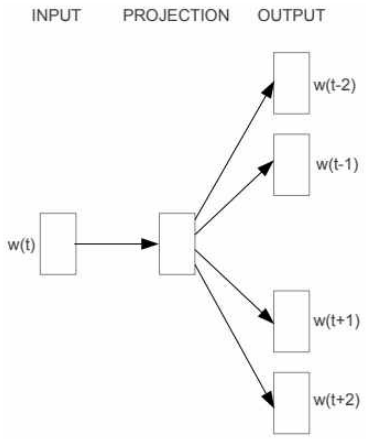


Fig. 1. The structure of skip-gram model.

Text는 단어 사전에 포함되지 않은 단어가 입력으로 주어지는 경우에도 입력 단어의 벡터 표현을 구할 수 있다는 장점을 가진다.

3.3 BERT

BERT(Bidirectional Encoder Representations from Transformers)[8]는 많은 자연어 응용 분야에서 좋은 성능을 내는 언어 모델이다. BERT 모델의 입력 표현은 token embedding, segment embedding, position embedding의 합으로 구성된다. token embedding은 WordPiece embedding[21]으로 표현되며, 자주 등장하는 sub-word는 그 자체가 단위가 되고, 자주 등장하지 않는 단어는 sub-word로 쪼개지게 된다. WordPiece 임베딩은 모든 언어에 적용 가능하며 sub-word 단위로 단어를 분절하므로 OOV 처리에 효과적이고 정확도 상승효과도 있다. segment embedding은 입력 데이터가 두 문장의 쌍으로 구성된 경우, 해당 토큰이 어느 문장에 포함되는지를 나타낸다. position embedding은 입력 토큰의 위치 정보를 줌으로써 동일 단어가 임베딩 되더라도 단어의 발생 위치와 문맥에 따라 서로 다른 임베딩 결과를 가지게 한다.

BERT는 transformer[22]모델의 Encoder 부분을 사용하여 양방향 모델을 구성한다. Transformer의 Encoder는 self-attention을 통해 각 단어에 대한 문맥적 단어 표현을 생성한다. BERT의 학습은 다음 두 가지 작업을 수행함으로써 진행된다.

- 입력의 일부분을 [MASK] 처리한 뒤 이를 예측하는 MLM(Masked Language Model) 작업
- 뒤이어 오는 문장이 실제로 이어지는 문장인지 예측하는 NSP(Next Sentence Prediction) 작업

MLM 학습에서는 입력 시퀀스에서 15%에 해당하는 무작위 단어를 몇 가지 규칙에 따라 [MASK] 토큰으로 대체한다. 이후 [MASK] 토큰이 있는 위치의 원본 단어를 예측하는 방법으로 학습을 수행한다. NSP 학습에서는 코퍼스에서 두 문장을 이어 붙인 뒤, 두 문장이 서로 연결되던 문장인지를 맞추는 이진 예측을 통해 모델을 학습한다. BERT를 이용한 자연어처리는 Encoder가 입력 문장들을 임베딩 하여 언어를 모델링하는 언어 모델링 구조 과정과 이를 fine-tuning하여 여러 자연어 처리 Task를 수행하는 과정으로 이루어진다.

4. 텍스트 임베딩과 이상 패턴 탐지를 이용한 신규 주제 발생 탐지

텍스트 데이터 스트림에서 신규 주제 발생 탐지를 위해 먼저 텍스트 임베딩 모델을 사용하여 데이터 스트림에 있는 각 문서를 벡터로 변환한다. 같은 주제에 관련된 문서들이 근접 거리에 매핑되도록 텍스트 임베딩에 의해 벡터로 변환함으로써 같은 주제 내 문서 밀집도를 높일 수 있을 것으로 기대된다. 변환된 데이터 스트림에서 논문 [3]에서 제안된 이상 패턴 탐지 방법을 적용하여 정상 데이터 분포에서 벗어나는 이상 패턴 발생을 탐지함으로써 정상 데이터 스트림에서 존재하지 않았던 신규 주제의 발생을 탐지할 수 있다. 본 논문에서는 텍스트 스트리밍 데이터에 포함되어 있는 텍스트 데이터들에 대한 벡터 표현을 구하기 위해 세 가지 임베딩 방법 BOW, Word2Vec, BERT를 사용하여 성능을 비교하였다.

4.1 텍스트 임베딩

[BOW] 스트리밍 데이터에서의 Tf-Idf 계산은 다음 과정을 통해 수행하였다. 먼저, 주어진 학습 데이터를 사용하여 단어들의 역문서 빈도를 계산한다. 이후 텍스트 스트림의 각 텍스트 데이터에 대한 임베딩 과정에서는 각 단어들의 단어 빈도에 학습데이터를 이용해 계산한 역문서 빈도를 곱해줌으로써 Tf-Idf 벡터를 구하였다.

[Word2Vec 임베딩] 단어들의 Word2Vec 표현을 구하기 위해 [23]에 공개되어 있는 사전 학습된 Word2Vec 모델을 사용하였다. 모델을 사용하여 텍스트에 포함되어 있는 단어들의 벡터들을 모두 구하고 모든 단어 벡터들의 평균을 계산한 값을 텍스트의 벡터로 사용하였다. 단어 벡터 계산 과정에서 Word2Vec 모델 학습 단어 사전에 포함되어 있지 않는 단어가 발생하는 경우, 해당 단어는 텍스트 벡터 계산 과정에서 제외하였다. 생성되는 텍스트의 벡터 표현의 차원 수는 사전 학습된 Word2Vec 모델의 임베딩 차원 수인 300이 된다.

[BERT 임베딩] 텍스트를 문단 단위로 나눈 뒤,

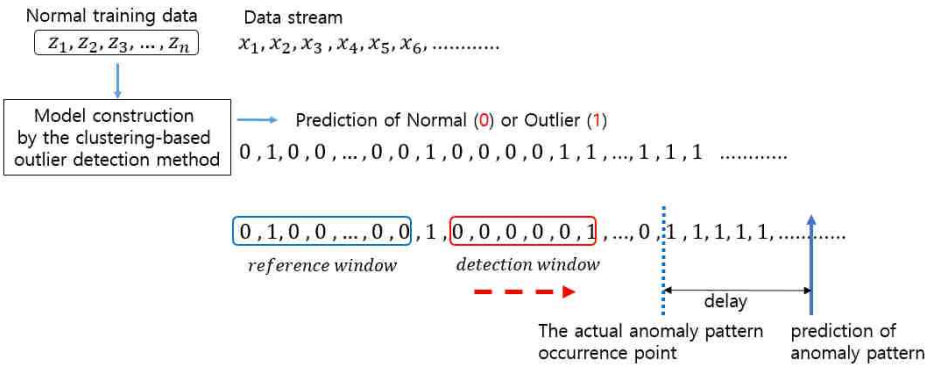


Fig. 2. Anomaly pattern detection in streaming data[3].

사전 학습된 BERT 모델[8]을 사용하여 각 문단 별 임베딩 표현을 구한 후 문단 벡터의 평균을 계산하였다. BERT 모델을 사용하여 임베딩을 수행하는 경우 모델의 출력 벡터는 입력 문장의 길이가 n일 때, 크기가 (n, dim)이 된다. 이를 (1, dim) 형태의 고정 길이 벡터로 변환하기 위해 bert-as-service[24]의 REDUCE_MEAN 옵션을 사용하였다. REDUCE_MEAN 옵션은 (n, dim)형태의 벡터 평균을 취해줌으로써 (1, dim) 형태의 벡터로 변환할 수 있다. 문서의 벡터 표현은 사전 학습된 모델의 은닉층 크기인 768이 된다. 사전 학습된 BERT 모델의 파라미터는 다음 Table 1과 같다.

4.2 스트리밍 데이터에서 이상 패턴 탐지

이상치 탐지는 개별적인 데이터 샘플에 대해 정상 데이터 분포를 벗어나는 이상치인지 정상인지 예측한다. 반면에 스트리밍 데이터에서의 이상 패턴 탐지는 정상 데이터 패턴을 벗어나는 이상치 발생이 갑자기 증가하게 되는 발생 지점 탐지를 목적으로 한다[3]. 먼저 정상 데이터로 구성된 학습 데이터가 주어졌다는 가정 하에 학습데이터를 이용하여 클러스터링 기반 앙상블 모델에 의한 이상치 탐지 모델을 구성한다. 주어진 정상 학습 데이터를 t개의 그룹으로 나눈 후, 각 그룹마다 k-means 클러스터링을 진행하여 각 그룹의 데이터가 나타내는 정상 영역을 포함하는 k개의 hypersphere를 구성한다. 테스트 데이터가 주어졌을 때, 그로부터 가장 가까운 클러스터의 중심을 찾고, 테스트 데이터가 해당 클러스터의 반경 밖에 있을 때 정상 영역을 벗어난다고 판단하게 된다. 테스트 데이터가 t개의 그룹 각각에 대한 클러스터링에서 정상 영역 밖에 있게 될 때, 테스트 데이터를 이상치로 예측한다. 이후 실시간으로 발생하는 데이터 스트림에 있는 각 데이터 샘플에 대해 이상치 탐지 모델을 적용하여 데이터 스트림을 0과 1의 이진값 스트림으로 변환할 수 있다. 0은 정상 데이터 예측을, 1은 이상치 예측을 나타낸다. Fig. 2에서처럼 이진값 스트림의 시작 부분에 참조 윈도우를 고정하고 탐지 윈도우는 새로운 데이터 샘플이 도착함에 따라 앞으로 전진하면서 탐지 윈도우 내에 이상 패턴 발생을 탐지한다. 두 윈도우에서의 이항 분포의 모비율이 같은지에 대한 가설 검정을 이용하여 탐지 윈도우에서 1의 빈도가 급격히 커지게 되는 시점을 탐지한다.

Table 1. Parameter setting of BERT model 모델

Parameter	Value
Hidden Size	768
Hidden Layers	12
Attention Heads	12
max position embedding	512
Dropout rate	0.1

5. 실험결과

5.1 실험 환경

성능 비교를 위하여 BBC News[25] 데이터셋을 사용하였다[1]. BBC News 데이터는 2004년부터 2005년까지 BBC 뉴스 웹사이트에서 다섯 개의 카테고리 business, entertainment, politics, sport, tech와 관련

1) http://mlg.ucd.ie/datasets/bbc.html에서 데이터를 다운받음

된 2,225개의 뉴스 데이터로 구성된 텍스트 데이터셀이다. 실험에서는 하나의 카테고리를 새로 발생하는 주제가 되도록 설정하여 탐지 성능을 측정하였다. 정상데이터로 구성되는 학습데이터는 신규 주제 카테고리를 제외한 나머지 네 개의 각 카테고리에 속하는 텍스트 데이터들의 60%를 랜덤하게 선택하여 구성하였다. 학습데이터에 포함되지 않은 각 카테고리의 나머지 40% 데이터와 함께 신규 카테고리에 속하는 텍스트 중에서 선택한 20%의 데이터를 합하여 테스트 데이터 스트림을 구성하였다. 테스트 데이터 스트림의 총 길이의 중간까지는 신규 주제가 포함되지 않게 기존 네 개의 카테고리의 데이터들만 나오게 하였고, 나머지 데이터 스트림 부분에서는 기존 카테고리와 신규 카테고리의 데이터가 랜덤하게 섞이게 하였다. 신규 카테고리의 데이터가 처음 나타나는 지점을 이상 패턴 발생 시점으로 설정하였다.

테스트 데이터 스트림에서 이상 패턴 탐지 성능은 논문 [3]에서와 같이 precision, recall, F1, delay를 사용하여 측정하였다. TP(True positive)는 실제 이상 패턴 존재 시점 이후에 이상 패턴의 발생을 예측할 경우이며, FP(False Positive)는 실제 이상 패턴 존재 시점 이전에 이상 패턴의 발생을 탐지할 경우, FN(False Negative)는 테스트 데이터 스트림에서 이상 패턴의 발생을 탐지하지 못할 경우를 나타낸다. 랜덤하게 데이터를 섞어서 학습 데이터와 테스트 데이터 스트림 구성을 10번 반복하여 실험을 수행하고 누적된 TP, FP, FN으로부터 precision, recall, F1을 식(2)로 구할 수 있다.

$$\text{precision} = TP/(TP+FP),\quad \text{recall} = TP/(TP+FN),$$
$$F1 = 2*\text{precision}*\text{recall}/(\text{precision}+\text{recall}) \qquad (2)$$

Fig. 2에서 테스트 데이터 스트림에서 실제 이상 패턴 존재 시점에서 이상 패턴 발생 예측 지점까지의 거리를 나타내는 delay의 측정에 대한 예시를 나타내었다. 10번 반복 실험 중에서 TP가 얻어진 경우에만 Fig. 2에서와 같이 delay를 측정하여 평균값을 구하였다.

5.2 신규 주제 발생 탐지 성능 비교

Table 2에서는 BBC News 데이터에서 각 카테고리를 새로 발생하는 주제로 설정했을 때 탐지 성능을 비교한다. 이상치 탐지 모델 구성을 위한 k-means 클러스터링에서 클러스터의 수는 40으로, 이상 패턴 탐지를 위한 윈도우의 크기는 80, 가설 검정 유의 수준을 0.01로 설정하였다. 텍스트 임베딩 방법으로써 BOW, Word2Vec, BERT를 사용하였다. 실험 결과 임베딩 방법으로 BERT 모델을 사용한 경우가 탐지 성능이 가장 높았고 실제 신규 주제 발생 시점부터 발생 탐지 알림까지의 delay가 가장 짧았다. FSD 방법과의 성능비교를 위해 논문 [26]에서 설명된 방법을 사용하여, 문서들을 Tf-Idf 벡터로 나타내고 테스트 스트림의 각 데이터에 대해 가장 가까운 코사인 유사도를 가지는 데이터 샘플과의 거리를 이용하여 문서의 novelty score를 계산하였다.

$$novelty\ score(d) = 1 - \max_{d' \in D_t} \cos(d, d') \qquad (3)$$

논문 [26]에서 D_t는 문서 d가 도착한 시간 t 이전까지의 모든 문서들의 집합으로 설정되었으나, 새로운 주제 데이터를 탐지하지 못하고 놓쳤을 때 그 데이터가 이후 진행 과정에서 성능을 저하시키는 요인으로 작용하는 것을 실험으로 확인하여 본 실험에서는 D_t 대신에 정상 학습데이터를 사용하였다. novelty score가 주어진 임계값 이상일 때 새로운 주제의 문

Table 2. A performance comparison when one category is set as a new emerging topic

New category	FSD		Text embedding model					
			BOW		Word2Vec		BERT	
	F1	delay	F1	delay	F1	delay	F1	delay
Business	0.89	92.5	0.57	146	0.89	101.2	1	111
Entertainment	0.95	137.8	0.46	224.3	0.82	114.9	1	80.7
Politics	0.89	156.4	0	-	0	-	0.57	178.8
Sport	0.89	36.8	0	-	0.95	107.8	1	65.1
Tech	0.75	99.7	0.18	302	0.33	129.5	0.67	127.6
Average	0.87	104.6	0.24	224.1	0.60	113.3	0.85	112.6

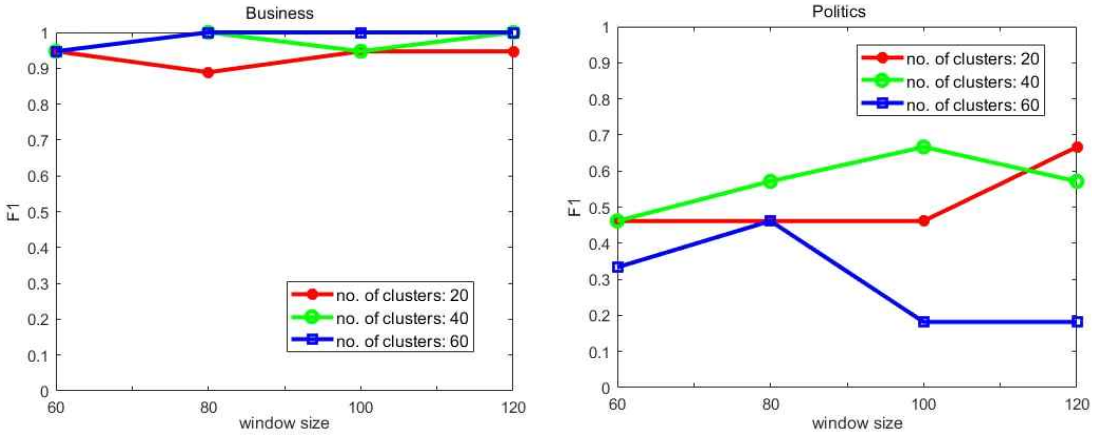
서로 선언하고 그 발생 시점을 신규 주제 발생 시점으로 하여 F1과 delay를 측정하였다. 임계값을 0.1부터 1까지 0.1 간격으로 설정하여 F1 성능이 가장 높은 경우를 구하고자 하였다. 그러나 모든 임계값 설정에서 F1 값은 0이었다. 따라서, 식 (2)의 novelty score를 모든 테스트 데이터에 대해 구한 후 신규 카테고리의 데이터와 기존 카테고리의 데이터의 novelty score들의 경계가 되는 구간을 구하고, 이 구간에서 0.01 간격으로 세분하여 임계값을 설정함으로써 성능을 테스트하였다. 모든 경우 중에서 가장 좋은 성능값을 Table 2에 나타내었다. FSD 방법은 임계값 설정을 매우 정밀하게 조정함으로써 높은 성능을 얻을 수 있으나, 실제 적용에서는 이와 같은 방식의 임계값 설정이 어렵다는 단점이 있다.

Table 3에서는 실험 설정을 달리하여 두 개의 카테고리의 데이터가 신규로 발생한다고 할 때 탐지 성능을 보여준다. Table 2에서는 BERT에 의한 임베딩을 이용할 때도 Politics와 Tech 클래스가 새로 나타나는 주제일 때 탐지가 어려웠던 반면, Table 3에서는 BERT를 이용할 때 평균적으로 0.92의 높은 F1의 값을 보였다. 두 개의 카테고리를 신규 주제로 설정한 실험에서 성능이 향상된 원인은 높은 탐지 능력을 갖는 카테고리의 영향과 신규 카테고리의 데이터 비중이 한 개의 카테고리를 신규 주제로 설정했을 때보다 커지게 된 때문이라고 여겨진다.

Table 2와 Table 3에서는 이상치 탐지 모델 구성을 위한 k-means clustering에서 클러스터 수는 40이고 이상 패턴 탐지를 위한 윈도우의 크기는 80으로 설정하였다. 매개변수들의 값에 대한 성능변화를 보기 위해 새로 나타나는 카테고리가 Business일 때와 Politics일 때 다양한 매개변수 값들의 사용에 따른 F1값을 Fig. 3에 나타내었다. Business가 새로 나타나는 카테고리일 때는 테스트된 매개변수 범위 내에서 전반적으로 안정적인 성능을 얻을 수 있음을 보여

Table 3. A performance comparison when two categories are set as new emerging topics

New category	Text embedding model					
	BOW		Word2Vec		BERT	
	F1	delay	F1	delay	F1	delay
Business+Entertainment	0.89	134.1	0.89	62.5	1	37.7
Entertainment+Politics	0.57	220.5	0.95	142.2	0.95	59.9
Politics+Tech	0	-	0.33	89.5	0.82	93.6
Sport+Tech	0.46	162.7	0.95	71.1	0.89	36.6
Average	0.48	172.4	0.78	91.3	0.92	57.0



(a) (b)

Fig. 3. Performance comparison when different parameter values were used. (a) Business category was set as a newly emerging topic (b) Politics category was set as a newly emerging topic.

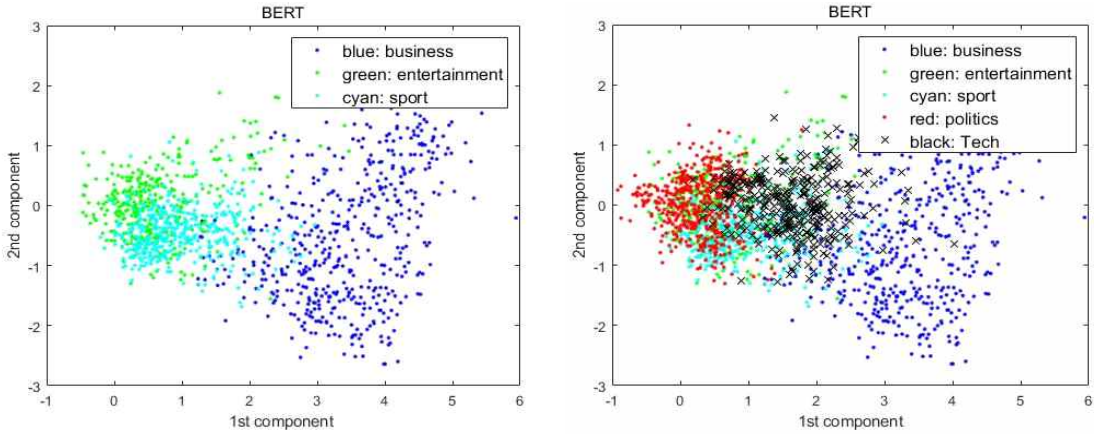


Fig. 4. Dimension reduction by PCA after BERT embedding (a) visualization for categories of business, entertainment, sport (b) visualization of five categories.

주나, Politics의 경우에는 클러스터 수가 작게 설정될 때 보다 나은 성능을 보였다.

5.3 차원 감소 방법 적용

텍스트 데이터에서는 포함되는 단어들이 기본적인 속성을 구성하기 때문에 BOW 방식에서와 같이 고차원 벡터로 표현된다. 고차원 데이터 분석에서는 특징 추출을 통해 차원을 낮추는 차원 감소 방법을 통해 데이터 분포를 시각화하거나 분석 성능을 향상시킬 수도 있다. 대표적인 무감독 차원 감소 방법인 PCA(Principal Component Analysis)를 적용하여 카테고리의 분포를 시각화하여 신규 주제 발생 탐지 성능 결과를 분석해보았다.

Fig. 4(a)와 (b)에서는 BERT에 의해 768차원으로 임베딩 된 데이터들을 PCA에 의해 2차원으로 차원 감소시켜 나타내었다. Fig. 4(a)에서는 Politics와 Tech를 제외한 세 개의 카테고리를 나타내고 (b)에서는 다섯 개 카테고리를 모두 나타냈다. 이상 패턴 탐지 성능이 가장 낮았던 Politics와 Tech 클래스가 다른 클래스들과 오버랩이 큰 것을 볼 수 있다. 기존의 카테고리와 오버랩이 큰 분포를 가지고 새로운 주제의 클래스가 발생할 때 탐지가 어려울 수 있음을 알 수 있다.

6. 결 론

본 논문에서는 텍스트 스트리밍 데이터에서 텍스트 임베딩과 이상 패턴 탐지를 이용한 신규 주제 발생 탐지 방법을 제안하였다. 유사한 의미의 단어나 문장들이 근접 거리에 위치하도록 단어나 문장을 다차원 공간에 벡터로 표현하는 임베딩 방법을 이용함으로써 문서 주제에 따른 군집화 성능을 높일 수 있다. 따라서 임베딩에 의해 벡터로 변환된 데이터들의 스트림에서 이상 패턴 탐지 방법의 적용은 학습데이터에 포함되지 않은 주제의 발생을 효과적으로 탐지할 수 있다. 실험에서는 BOW, Word2Vec, BERT의 텍스트 임베딩 방법을 사용하여 성능을 비교하였다. 실험 결과는 BERT 모델을 사용했을 때 신규 주제 발생 탐지 성능이 높음을 보여주었다. 본 논문에서의 신규 주제 발생 탐지는 정상 데이터를 사용하여 이상치 탐지 모델을 학습하고, 이를 사용하여 이상 패턴 발생 탐지를 수행하여 텍스트 스트림에서 기존에 등장하지 않던 신규 주제가 발생하는 시점을 탐지하는 것에 중점을 두었다. 향후 연구로 신규 주제의 발생을 탐지한 시점에 탐지 윈도우 내의 이상치 문서들을 사용하여 키워드 추출, 문서 요약 등의 기법을 사용하여 발생한 신규 주제 분석을 수행할 수 있을 것으로 기대된다.

REFERENCE

[1] C. Aggarwal, *Outlier Analysis*, Springer, Switzerland, 2017.

[2] C. Park, “Outlier and Anomaly Pattern Detection on Data Streams,” *The Journal of Supercomputing*, Vol. 75, No. 9, pp. 6118-6128, 2019.

[3] T. Kim and C. Park, "Anomaly Pattern Detection for Streaming Data," *Expert Systems with Applications*, Vol. 149, pp. 1-8, 2020.

[4] C. Park and T. Kim, "Energy Theft Detection in Advanced Metering Infrastructure Based on Anomaly Pattern Detection," *Energies*, Vol. 13, No. 15, pp. 1-10, 2020.

[5] J. Allan, R. Papka, and V. Lavrenko, "On-line New Event Detection and Tracking," *Proceeding of International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 37-45, 1998.

[6] J. Allan, "Introduction to Topic Detection and Tracking," *In Topic Detection and Tracking: Event-based Information Organization*, Vol. 12, pp. 1-16. 2002.

[7] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and Their Compositionality," *Proceeding of International Conference on Neural Information Processing Systems*, Vol. 2, pp. 3111-3119, 2013.

[8] J. Devlin, M.W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1, pp. 4171-4186, 2019.

[9] S. Choi, *New Topic Occurrence Detection Using Text Embedding Model in Text Streaming Data*, Master's Thesis of Chungnam National University, 2020.

[10] S. Petrovi´c, M. Osborne, and V. Lavrenko, "Streaming First Story Detection with Application to Twitter," *Proceeding of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 181-189, 2010.

[11] S. Moran, R. McCreadie, C. Macdonald, and I. Ounis, "Enhancing First Story Detection Using Word Embeddings," *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 821-824, 2016.

[12] N. Panagiotou, C. Akkaya, K. Tsioutsiouliklis, V. Kalogeraki, and D. Gunopulos, "First Story Detection Using Entities and Relations," *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers,* pp. 3237-3244, 2016.

[13] E. Lee and P. Kim, "A Method for Short Text Classification Using SNS Feature Information Based on Markov Logic Network," *Journal of Korea Multimedia Society*, Vol. 20, No. 7, pp. 1065-1072, 2017.

[14] M. Mathioudakis and N. Koudas, "Twitter Monitor: Trend Detection Over the Twitter Stream," *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*, pp. 1155-1158, 2010.

[15] H.J. Choi and C.H. Park, "Emerging Topic Detection in Twitter Stream Based on High Utility Pattern Mining," *Expert Systems with Applications*, Vol. 115, pp. 27-36, 2019.

[16] S. Phuvipadawat and T. Murata, "Breaking News Detection and Tracking in Twitter," *Proceedings of IEEE/WIC/ACM International Conference Web Intelligence and Intelligent Agent Technology*, pp. 120-123, 2010.

[17] D. Quercia, H. Askham, and J. Crowcroft, "Tweetlda: Supervised Topic Classification and Link Prediction in Twitter," *Proceedings of 4th Annual ACM Web Science Conference*, pp. 247-250, 2012.

[18] U. Erra, S. Senatore, and G. Caggianese. "Approximate TF-IDF Based on Topic Extraction from Massive Message Stream Using the GPU," *Information Sciences*, Vol. 292, pp. 143-161, 2015.

[19] G. Salton, E.A. Fox, and H. Wu, "Extended Boolean Information Retrieval," *Communica-*

tions of the ACM, Vol. 26, No. 11, pp. 1022-1036, 1983.

[20] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching Word Vectors with Subword Information," *Transactions of the Association for Computational Linguistics*, Vol. 5, No. 1, pp. 135-146, 2017.

[21] Y. Wu, M. Schuster, Z. Chen, Q. Le, M. Norouzi, W. Macherey, et al., "Google's Neural Machine Translation System: Bridging the Gap Between Human and Machine Translation," *arXiv Preprint arXiv:1609.08144*, 2016.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, et al., "Attention is All You Need," *Advances in Neural Information Processing Systems*, pp. 5998-6008, 2017.

[23] Word2vec(2013), https://code.google.com/archive/p/word2vec/ (accessed February 20, 2019).

[24] Bert-as-service(2018), https://github.com/hanxiao/bert-as-service (accessed February 20, 2019).

[25] D. Greene and P. Cunningham, "Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering," *Proceedings of International Conference on Machine Learning*, pp. 377-384, 2006.

[26] S. Petrovic, M. Osborne, and V. Lavrenko, "Using Paraphrases for Improving First Story Detection in News and Twitter," *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 338-346, 2012.

최 세 목

2018년 충남대학교 컴퓨터공학과 졸업(공학사)
2020년 충남대학교 컴퓨터공학과 석사졸업(공학석사)
관심분야 : 자연어처리, 데이터 마이닝, 이상치 탐지, 인공지능 등

박 정 희

1998년 연세대학교 수학과 (박사)
2004년 University of Minnesota, Computer Science & Engineering(박사)
2005년~현재 충남대학교 컴퓨터공학과 교수

관심분야 : Data Mining, 기계학습