【발명의 설명】

【발명의 명칭】

도메인 감성 사전과 워드 임베딩 기법을 결합한 감성 분석 방법 및 시스템{SENTIMENT ANALYSIS METHOD AND SYSTEM COMBINING DOMAIN SENTIMENT DICTIONARY AND WORD EMBEDDING TECHNIQUE}

【기술분야】

<0001> 아래의 설명은 감성 분석 기술에 관한 것이다.

<0002>

【발명의 배경이 되는 기술】

<0003> 감성 분석은 텍스트에 내포되어 있는 작성자의 의견, 성향 등을 분석하는 기술이다. 최근 다양한 도메인을 대상으로 수요가 증가되고 있다. 예를 들면, 경제, 사회에 대한 긍정 또는 부정적인 의견이나, 영화, 드라마에 대한 시청자 평가나, 제품, 서비스에 대한 소비자 평가 등이 해당된다.

<0004> 감성 분석을 위해 어휘 기반 분석 방법 또는 딥러닝 기반 분석 방법이 이용되고 있다. 어휘 기반 분석 방법은 수작업, 통계 또는 워드 임베딩을 이용한 말뭉치를 기반으로 감성 사전을 구축하여 감성을 분석한다. 딥러닝 기반의 분석 방법은 인공신경망, BERT 등을 이용하여 감성을 분석한다.

<0005> 그러나, 종래의 기술은 동일한 단어라도 도메인에 따라 다른 감성을 가질 수 있는 것을 고려하고 있지는 못하고 있다.

<0006>

【발명의 내용】

【해결하고자 하는 과제】

<0007> 말뭉치를 기반으로 도메인 특성이 반영된 도메인 감성 사전을 구축하는 방법 및 시스템을 제공할 수 있다.

<0008> 도메인 감성 사전과 워드 임베딩이 결합된 감성 분석 모델을 이용하여 새로운 말뭉치로부터 도메인에 따른 감성을 분류하는 방법 및 시스템을 제공할 수 있다.

<0009>

【과제의 해결 수단】

<0010> 감성 분석 시스템에 의해 수행되는 감성 분석 방법은, 말뭉치를 기반으로 도메인 특성이 반영된 도메인 감성 사전을 구축하는 단계; 및 상기 구축된 도메인 감성 사전과 워드 임베딩이 결합된 감성 분석 모델을 이용하여 새로운 말뭉치로부터 도메인에 따른 감성을 분류하는 단계를 포함하고, 상기 감성 분석 모델은, 말뭉치에 대한 학습 데이터를 이용하여 각 어휘의 도메인 감성 스코어와 각 어휘의 유사도가 반영되어 도메인에 따라 감성이 분류되도록 학습된 것일 수 있다.

<0011> 상기 도메인 감성 사전을 구축하는 단계는, 라벨링된 데이터 셋의 라벨 값을 참조하여 텍스트 데이터(document)에 대한 부정 또는 긍정을 포함하는 문맥 정보를 파악하고, 상기 파악된 문맥 정보를 이용하여 각 어휘에 대한 감성 점수를 부여하는 단계를 포함할 수 있다.

<0012> 상기 도메인 감성 사전을 구축하는 단계는, 상기 각 어휘에 대한 감성 점수

를 시그모이드 함수를 사용하여 정규화하고, 상기 각 어휘에 대한 품사를 지정하고, 상기 지정된 각 어휘에 대한 품사에 따라 어휘별 감성 점수를 계산하는 단계를 포함할 수 있다.

<0013> 상기 도메인 감성 사전을 구축하는 단계는, 상기 구축된 도메인 감성 사전의 점수와 범용 감성 사전의 점수를 비교하여 상기 구축된 도메인 감성 사전을 평가하는 단계를 포함할 수 있다.

<0014> 상기 분류하는 단계는, 상기 구축된 도메인 감성 사전을 감성 스코어 임베딩 레이어로 변환하고, 상기 변환된 감성 스코어 임베딩 레이어와 워드 임베딩 레이어가 결합된 감성 분석 모델을 생성하고, 말뭉치에 대한 학습 데이터를 이용하여 상기 생성된 감성 분석 모델을 학습시키는 단계를 포함할 수 있다.

<0015> 상기 분류하는 단계는, 상기 감성 스코어 임베딩 레이어를 통한 감성 스코어 임베딩을 통해 도메인 감성 사전을 벡터로 표현하고, 상기 워드 임베딩 레이어를 통한 워드 임베딩을 통해 각 어휘의 유사도를 반영하기 위한 벡터로 표현하는 단계를 포함할 수 있다.

<0016> 상기 말뭉치에 대한 학습 데이터는, 상기 감성 분석 모델의 학습을 위해 데이터 전처리가 수행된 것일 수 있다.

<0017> 상기 데이터 전처리는, 텍스트 데이터에서 특수 문자를 포함하는 불필요한 문자를 제거하는 데이터 정제 과정; 상기 불필요한 문자가 제거된 텍스트 데이터에서 불용어를 제거하는 불용어 제거 과정; 상기 불용어가 제거된 텍스트 데이터에서 형태소 정규화를 통해 어휘를 추출하는 어간 추출 과정; 상기 형태 정규화를 수행

한 어휘마다 번호를 부여하여 토큰화된 어휘를 생성하는 토큰화 과정; 상기 토큰화된 어휘를 기반으로 감성 분석에 사용할 어휘를 토큰 번호로 변환하는 인코딩 과정; 및 상기 감성 분석에 사용할 어휘를 동일한 크기의 벡터로 변환하는 패딩 과정을 수행할 수 있다.

<0018> 감성 분석 시스템은, 말뭉치를 기반으로 도메인 특성이 반영된 도메인 감성 사전을 구축하는 사전 구축부; 및 상기 구축된 도메인 감성 사전과 워드 임베딩이 결합된 감성 분석 모델을 이용하여 새로운 말뭉치로부터 도메인에 따른 감성을 분류하는 감성 분류부를 포함하고, 상기 감성 분석 모델은, 말뭉치에 대한 학습 데이터를 이용하여 각 어휘의 도메인 감성 스코어와 각 어휘의 유사도가 반영되어 도메인에 따라 감성이 분류되도록 학습된 것일 수 있다.

<0019>

【발명의 효과】

<0020> 도메인 감성 사전과 워드 임베딩의 결합을 통해 생성된 감성 분석 모델을 이용하여 감성 분석의 성능을 향상시킬 수 있다.

<0021> 도메인 감성 사전을 감성 스코어 임베딩 레이어로 변환하고, 워드 임베딩 레이어를 통해 텍스트의 각 단어를 벡터로 표현하여 도메인에 따른 감성을 분류할 수 있다.

<0022>

【도면의 간단한 설명】

<0023> 도 1은 일 실시예에 따른 감성 분석 시스템의 구성을 설명하기 위한 블록도

이다.

도 2는 일 실시예에 따른 감성 분석 시스템에서 감성 분석 방법을 설명하기 위한 흐름도이다.

도 3은 일 실시예에 있어서, 문맥 정보를 파악하는 동작을 설명하기 위한 예이다.

도 4는 일 실시예에 있어서, 데이터 전처리 동작을 설명하기 위한 도면이다.

도 5 내지 도 7은 일 실시예에 있어서, 도메인 감성 사전을 구축하는 동작을 설명하기 위한 도면이다.

도 8은 일 실시예에 있어서, 도메인 감성 사전을 평가하는 동작을 설명하기 위한 예이다.

도 9는 일 실시에에 있어서, 감성 분석 모델의 구조를 설명하기 위한 도면이다.

도 10은 일 실시예에 있어서, 워드 임베딩을 설명하기 위한 도면이다.

도 11은 일 실시에에 있어서, 감성 감성 스코어 임베딩을 설명하기 위한 도면이다.

도 12는 일 실시예에 있어서, 임베딩 레이어를 결합하는 동작을 설명하기 위한 도면이다.

【발명을 실시하기 위한 구체적인 내용】

<0024> 이하, 실시예를 첨부한 도면을 참조하여 상세히 설명한다.

<0025>

<0026> 도 1은 일 실시예에 따른 감성 분석 시스템의 구성을 설명하기 위한 블록도이고, 도 2는 일 실시예에 따른 감성 분석 시스템에서 감성 분석 방법을 설명하기 위한 흐름도이다.

<0027> 감성 분석 시스템(100)의 프로세서는 사전 구축부(110) 및 감성 분류부(120)를 포함할 수 있다. 이러한 프로세서의 구성요소들은 감성 분석 시스템에 저장된 프로그램 코드가 제공하는 제어 명령에 따라 프로세서에 의해 수행되는 서로 다른 기능들(different functions)의 표현들일 수 있다. 프로세서 및 프로세서의 구성요소들은 도 2의 감성 분석 방법이 포함하는 단계들(210 내지 220)을 수행하도록 감성 분석 시스템을 제어할 수 있다. 이때, 프로세서 및 프로세서의 구성요소들은 메모리가 포함하는 운영체제의 코드와 적어도 하나의 프로그램의 코드에 따른 명령(instruction)을 실행하도록 구현될 수 있다.

<0028> 프로세서는 감성 분석 방법을 위한 프로그램의 파일에 저장된 프로그램 코드를 메모리에 로딩할 수 있다. 예를 들면, 감성 분석 시스템에서 프로그램이 실행되면, 프로세서는 운영체제의 제어에 따라 프로그램의 파일로부터 프로그램 코드를 메모리에 로딩하도록 감성 분석 시스템을 제어할 수 있다. 이때 사전 구축부(110) 및 감성 분류부(120) 각각은 메모리에 로딩된 프로그램 코드 중 대응하는 부분의 명령을 실행하여 이후 단계들(210 내지 220)을 실행하기 위한 프로세서의 서로 다른 기능적 표현들일 수 있다.

<0029> 단계(210)에서 사전 구축부(110)는 말뭉치를 기반으로 도메인 특성이 반영된 도메인 감성 사전을 구축할 수 있다. 사전 구축부(110)는 라벨링된 데이터 셋의

라벨 값을 참조하여 말뭉치(document)에 대한 부정 또는 긍정을 포함하는 문맥 정보를 파악하고, 파악된 문맥 정보를 이용하여 각 어휘에 대한 감성 점수를 부여할 수 있다. 도 3을 참고하면, 문맥 정보를 파악하는 동작을 설명하기 위한 예이다. 예를 들면, 사전 구축부(110)는 라벨링된 데이터 셋(예를 들면, 영화 리뷰)의 라벨 값을 참조하여 문맥이 부정(label=0)인지 긍정(label=1)인지 파악할 수 있다. 사전 구축부(110)는 말뭉치 기반 도메인 특성을 반영할 수 있다. 사전 구축부(110)는 텍스트 데이터(document)의 라벨 값에 따라 감성(부정=0, 긍정=1)을 분류하고 어휘에 대한 감성 점수를 부여할 수 있다. Label=0→부정→s_t=0, Label=1→긍정→s_t=1이 된다. 예를 들면, '이 영화를 계속 보고 있으니 잠이 온다.'라는 텍스트 데이터에서 Label=0이다. 라벨을 참고할 때, '잠이 온다'는 부정적인 의미로 쓰이고 있기 때문에 감성 점수는 음의 값을 가진다. 다른 예로서, '침대에 누워 있으니 나도 모르게 잠이 온다.'라는 텍스트 데이터에서 Label=1이다. 라벨을 참고할 때, '잠이 온다'는 긍정적인 의미로 쓰이고 있기 때문에 감성 점수는 양의 값을 가진다. 사전 구축부(110)는 말뭉치(전체 텍스트 데이터)에서 동일한 어휘에 대한 감성 점수를 계산하여 최종적으로 어휘에 대한 감성 점수를 부여할 수 있다.

<0030> 도 5를 참고하면, 도메인 감성 사전을 구축하는 동작에 대하여 설명하기로 한다. 사전 구축부(110)는 단어 정규화, 품사 태깅 및 감성 점수 계산을 통해 도메인 감성 사전을 구축할 수 있다. 도 7을 참고하면, 사전 구축부(110)는 도메인 감성 사전을 구축하기 위하여 어휘를 정규화할 수 있다. 사전 구축부(110)는 각 어휘에 대한 감성 점수를 시그모이드 함수를 사용하여 정규화할 수 있다. 사전 구

축부(110)는 딥러닝 네트워크와 결합 후 효율적인 학습을 위해 선형 출력을 비선형 출력으로 변환할 수 있다. 감성 점수 정규화는 다음과 같이 계산될 수 있다.

<0031> 수학식 1:

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

<0032>

<0033> 이때, x는 어휘에 대한 감성 점수, $\sigma(x)$는 감성 점수 정규화($0<\sigma(x)<1$)이다.

<0034> 사전 구축부(110)는 각 어휘에 대한 품사를 지정하고, 지정된 각 어휘에 대한 품사에 따라 어휘별 감성 점수를 계산할 수 있다. 다시 말해서, 사전 구축부(110)는 품사 태깅을 통해 형용사뿐만 아니라 품사를 지정하여 감성 점수를 계산할 수 있다. 사전 구축부(110)는 말뭉치에서 어휘의 각 감성 점수를 텍스트 데이터에서 $Score(w_p)$ 합으로 계산할 수 있다.

<0035> 감성 점수 계산은 도메인 감성 사전 수식에 의해 계산될 수 있다.

<0036> 수학식 2:

$$Score(w_p) = \sum_{t=1}^{n} (s_t / L_{tp})$$

<0037>

<0038> 여기서, w는 어휘, t는 텍스트, n는 텍스트 개수, p는 품사, L_{tp}는 텍스트에서 품사 개수, s_t는 텍스트 감성을 의미한다. 먼저 감성 점수를 부여할 단어 W_p의 품사(part-of-speech)를 지정하고, 각 텍스트 t에서 p의 개수 L_{tp}를 구한다. 이후 L_{tp}의 개수만큼 나눈 값 $1/L_{tp}$와 텍스트의 감성 s_t를 곱한다. 마지막으로 이 값을 모

든 텍스트에 적용하여 더한다.

<0039> 도 6을 참고하면, 영화 리뷰에서 '저렴하다'라는 어휘의 감성 점수를 계산하는 것을 예를 들어 설명하기로 한다. 영화 리뷰 말뭉치에서 '저렴하다'라는 어휘가 부정으로 라벨링된 텍스트에서 많이 쓰인다(음의 값으로 계산됨). 이에, 영화 리뷰 말뭉치에서 '저렴하다'라는 어휘의 최종 감성 점수가 -2.98174로 계산될 수 있다.

<0040> 도 8을 참고하면, 도메인 감성 사전을 평가하는 동작을 설명하기 위한 예이다. 도 8은 영화 리뷰의 어휘 감성 점수를 나타낸 예이다. 사전 구축부(110)는 구축된 도메인 감성 사전의 점수와 범용 감성 사전의 점수를 비교하여 구축된 도메인 감성 사전을 평가할 수 있다. 예를 들면, 사전 구축부(110)는 도메인 감성 사전의 점수와 KNU 한국어 감성 사전(범용 감성 사전)의 점수를 비교할 수 있다. 이때, 사전 구축부(110)는 도메인 특성에 맞게 감성 점수를 계산할 수 있다. 도메인 감성 사전 점수는 $S>0.5$일 경우 긍정, $S\leq 0.5$일 경우 부정으로 판단하고, KNU 한국어 감성 사전은 $S>0$일 경우 긍정, $S<0$일 경우 부정으로 판단한다.

<0041> 단계(220)에서 감성 분류부(120)는 구축된 도메인 감성 사전과 워드 임베딩이 결합된 감성 분석 모델을 이용하여 새로운 말뭉치로부터 도메인에 따른 감성을 분류할 수 있다. 이때, 감성 분석 모델은, 말뭉치에 대한 학습 데이터를 이용하여 각 어휘의 도메인 감성 스코어와 각 어휘의 유사도가 반영되어 도메인에 따라 감성이 분류되도록 학습된 것일 수 있다. 감성 분류부(120)는 구축된 도메인 감성 사전을 감성 스코어 임베딩 레이어로 변환하고, 변환된 감성 스코어 임베딩 레이어와

워드 임베딩 레이어가 결합된 감성 분석 모델을 생성하고, 말뭉치에 대한 학습 데이터를 이용하여 생성된 감성 분석 모델을 학습시킬 수 있다. 감성 분류부(120)는 학습된 감성 분석 모델을 이용하여 새로운 말뭉치에 대한 감성을 분류할 수 있다. 예를 들면, 감성 분류부(120)는 새로운 말뭉치에 대해 긍정 또는 부정의 감성을 분류할 수 있다.

<0042> 도 9를 참고하면, 감성 분석 모델의 구조를 설명하기 위한 도면이다. 감성 분석 모델(900)은 감성 스코어 임베딩 레이어(Sentiment score embedding layer), 워드 임베딩 레이어(Word embedding layer), 복수 개의 LSTM, 결합(concatenate) 및 복수 개의 완전 연결 레이어(Fully Connected layer)로 구성될 수 있다. 감성 분류부(120)는 감성 스코어 임베딩 레이어와 워드 임베딩 레이어를 결합한 감성 분석 모델(900)을 생성할 수 있다. 감성 분류부(120)는 감성 스코어 임베딩을 통해 도메인 감성 사전을 벡터로 표현할 수 있다. 감성 분류부(120)는 워드 임베딩을 통해 각 어휘의 유사도를 반영하기 위한 벡터로 표현할 수 있다. 이때, 유사도는 문맥에서 기준 어휘 주변의 어휘를 분석하여 어휘의 유사도가 판단될 수 있다. 감성 분류부(120)는 감성 스코어 임베딩 레이어와 워드 임베딩 레이어를 결합 후 완전 연결 레이어를 통해 감성을 분류할 수 있다.

<0043> 도 10을 참고하면, 워드 임베딩을 설명하기 위한 도면이다. 감성 분류부(120)는 워드 임베딩 레이어를 통한 워드 임베딩을 이용하여 어휘를 벡터로 표현할 수 있다. 감성 분류부(120)는 어휘를 정수로 인코딩할 수 있다. 예를 들면, '저렴하다'라는 어휘를 인코딩할 경우, 1918이라는 값이 도출될 수 있다(Encoder('

저렴하다') = 1918). 감성 분류부(120)는 룩업 테이블을 통해 어휘에 대한 벡터 값을 가져올 수 있다. 룩업 테이블에 어휘 인코딩 값(정수)가 입력 데이터로 입력될 수 있다. 룩업 테이블에 입력 데이터가 입력됨에 따라 어휘 인코딩 값을 키(key)로 사용하여 룩업 테이블에서 키에 해당하는 벡터 값이 결과 데이터로서 출력될 수 있다. 예를 들면, 인코딩된 '저렴하다'라는 값에 대한 룩업 테이블의 결과 데이터 [1.2, 0.7, 1.9, 1.5]가 출력될 수 있다(lookupTable(Encoder('저렴하다')) = [1.2, 0.7, 1.9, 1.5]). 실시예에서는 룩업 테이블의 차원을 300으로 설정하고, Word2Vec 알고리즘이 사용될 수 있다(도 10의 차원은 4).

<0044> 이때, 워드 임베딩을 위하여 Word2Vec, 유사도 반영 방법이 사용될 수 있다. Word2Vec란 어휘 간에 유사도를 반영하기 위한 임베딩 방법으로, 얕은 신경망 학습을 통해 두 어휘 간 유사도 점수 계산이 가능하다. 다시 말해서, 중심 단어와 주변 단어와의 관계 정보를 이용하여 벡터로 표현될 수 있다. Word2Vec는 딥러닝 네트워크에 사용할 수 있다. 유사도 반영 방법으로 CBOW(Continuous Bag Of Words), Skip-gram 등이 사용될 수 있다. CBOW는 주변 어휘를 통한 중간 예측 알고리즘이고, skip-gram은 중간 어휘를 통한 어휘 예측 알고리즘으로, CBOW 알고리즘의 역으로 동작된다.

<0045> 도 11을 참고하면, 감성 스코어 임베딩을 설명하기 위한 도면이다. 감성 분류부(120)는 딥러닝 학습에 사용하기 위해 감성 사전을 임베딩 레이어로 변환할 수 있다. 이때, 차원은 1로 설정될 수 있다.

<0046> 도 12를 참고하면, 임베딩 레이어를 결합하는 동작을 설명하기 위한 도면이

다. 결합(Concatenate)은 감성 스코어 임베딩 레이어와 워드 임베딩 레이어를 입력으로 하는 각 LSTM 레이어의 출력 연결이다. 각 임베딩 레이어를 직접 결합하는 것이 아니라 LSTM 레이어를 통과한 후 결합한다. 실시예에서 LSTM 레이어의 출력 개수 N은 256으로 설정된 것을 예를 들기로 한다. 완전 연결 레이어는 결합 레이어를 완전 연결 레이어에 완전 연결시켜 감성 분석 모델의 최종 출력(0<y<1)을 계산한다.

<0047> 도 4는 일 실시예에 있어서, 데이터 전처리 동작을 설명하기 위한 도면이다.

<0048> 감성 분석 모델의 학습을 위한 데이터 전처리가 수행될 수 있다. 데이터 전처리는 데이터 정제(410), 불용어 제거(420), 어간 추출(430), 토큰화(440), 인코딩(450) 및 패딩(450) 과정을 수행할 수 있다. 감성 분석 시스템은 텍스트 데이터에서 특수 문자를 포함하는 불필요한 문자를 제거하는 데이터 정제(410) 과정을 수행할 수 있다. 감성 분석 시스템은 불필요한 문자가 제거된 말뭉치에서 불용어를 제거하는 불용어 제거(420) 과정을 수행할 수 있다. 여기서 불용어란 "의", "가", "이", "은", "들", "는", "과", "도", "를", "으로", "에" 등과 같이 인터넷 검색 시 검색 용어로 사용하지 않는 단어를 의미한다. 관사, 전치사, 조사, 접속사 등 검색 색인 단어로 의미가 없는 단어를 의미한다. 감성 분석 시스템은 불용어가 제거된 말뭉치에서 형태소 정규화를 통해 어휘를 추출하는 어간 추출(430) 과정을 수행할 수 있다. 감성 분석 시스템은 형태 정규화를 수행한 어휘마다 번호를 부여하여 토큰화된 어휘를 생성하는 토큰화(440) 과정을 수행할 수 있다. 감성 분석 시스템은 토큰화된 어휘를 기반으로 감성 분석에 사용할 어휘를 토큰 번호로 변환하

는 인코딩(450) 과정을 수행할 수 있다. 감성 분석 시스템은 감성 분석에 사용할 어휘를 동일한 크기의 벡터로 변환하는 패딩(460) 과정을 수행할 수 있다.

<0049> 이와 같이, 감성 분석 모델을 학습하기 위한 학습 데이터가 전처리될 수 있다. 감성 분석 시스템(100)은 도메인 감성 사전과 워드 임베딩이 결합된 감성 분석 모델에 전처리된 학습 데이터를 입력받을 수 있다. 감성 분석 시스템(100)은 전처리된 학습 데이터를 이용하여 도메인 감성 사전과 워드 임베딩이 결합된 감성 분석 모델을 학습시킬 수 있다.

<0050> 이상에서 설명된 장치는 하드웨어 구성요소, 소프트웨어 구성요소, 및/또는 하드웨어 구성요소 및 소프트웨어 구성요소의 조합으로 구현될 수 있다. 예를 들어, 실시예들에서 설명된 장치 및 구성요소는, 예를 들어, 프로세서, 콘트롤러, ALU(arithmetic logic unit), 디지털 신호 프로세서(digital signal processor), 마이크로컴퓨터, FPGA(field programmable gate array), PLU(programmable logic unit), 마이크로프로세서, 또는 명령(instruction)을 실행하고 응답할 수 있는 다른 어떠한 장치와 같이, 하나 이상의 범용 컴퓨터 또는 특수 목적 컴퓨터를 이용하여 구현될 수 있다. 처리 장치는 운영 체제(OS) 및 상기 운영 체제 상에서 수행되는 하나 이상의 소프트웨어 애플리케이션을 수행할 수 있다. 또한, 처리 장치는 소프트웨어의 실행에 응답하여, 데이터를 접근, 저장, 조작, 처리 및 생성할 수도 있다. 이해의 편의를 위하여, 처리 장치는 하나가 사용되는 것으로 설명된 경우도 있지만, 해당 기술분야에서 통상의 지식을 가진 자는, 처리 장치가 복수 개의 처리 요소(processing element) 및/또는 복수 유형의 처리 요소를 포함할 수 있음을 알

수 있다. 예를 들어, 처리 장치는 복수 개의 프로세서 또는 하나의 프로세서 및 하나의 콘트롤러를 포함할 수 있다. 또한, 병렬 프로세서(parallel processor)와 같은, 다른 처리 구성(processing configuration)도 가능하다.

<0051> 소프트웨어는 컴퓨터 프로그램(computer program), 코드(code), 명령(instruction), 또는 이들 중 하나 이상의 조합을 포함할 수 있으며, 원하는 대로 동작하도록 처리 장치를 구성하거나 독립적으로 또는 결합적으로(collectively) 처리 장치를 명령할 수 있다. 소프트웨어 및/또는 데이터는, 처리 장치에 의하여 해석되거나 처리 장치에 명령 또는 데이터를 제공하기 위하여, 어떤 유형의 기계, 구성요소(component), 물리적 장치, 가상 장치(virtual equipment), 컴퓨터 저장 매체 또는 장치에 구체화(embody)될 수 있다. 소프트웨어는 네트워크로 연결된 컴퓨터 시스템 상에 분산되어서, 분산된 방법으로 저장되거나 실행될 수도 있다. 소프트웨어 및 데이터는 하나 이상의 컴퓨터 판독 가능 기록 매체에 저장될 수 있다.

<0052> 실시예에 따른 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 상기 매체에 기록되는 프로그램 명령은 실시예를 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(magnetic media), CD-ROM, DVD와 같은 광기록 매체(optical media), 플롭티컬 디스크(floptical disk)와 같은 자기-

광 매체(magneto-optical media), 및 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다.

<0053> 이상과 같이 실시예들이 비록 한정된 실시예와 도면에 의해 설명되었으나, 해당 기술분야에서 통상의 지식을 가진 자라면 상기의 기재로부터 다양한 수정 및 변형이 가능하다. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로 등의 구성요소들이 설명된 방법과 다른 형태로 결합 또는 조합되거나, 다른 구성요소 또는 균등물에 의하여 대치되거나 치환되더라도 적절한 결과가 달성될 수 있다.

<0054> 그러므로, 다른 구현들, 다른 실시예들 및 특허청구범위와 균등한 것들도 후술하는 특허청구범위의 범위에 속한다.

<0055>

【청구범위】

【청구항 1】

감성 분석 시스템에 의해 수행되는 감성 분석 방법에 있어서,

말뭉치를 기반으로 도메인 특성이 반영된 도메인 감성 사전을 구축하는 단계; 및

상기 구축된 도메인 감성 사전과 워드 임베딩이 결합된 감성 분석 모델을 이용하여 새로운 말뭉치로부터 도메인에 따른 감성을 분류하는 단계

를 포함하고,

상기 감성 분석 모델은, 말뭉치에 대한 학습 데이터를 이용하여 각 어휘의 도메인 감성 스코어와 각 어휘의 유사도가 반영되어 도메인에 따라 감성이 분류되도록 학습된 것을 특징으로 하는 감성 분석 방법.

【청구항 2】

제1항에 있어서,

상기 도메인 감성 사전을 구축하는 단계는,

라벨링된 데이터 셋의 라벨 값을 참조하여 텍스트 데이터(document)에 대한 부정 또는 긍정을 포함하는 문맥 정보를 파악하고, 상기 파악된 문맥 정보를 이용하여 각 어휘에 대한 감성 점수를 부여하는 단계

를 포함하는 감성 분석 방법.

【청구항 3】

제2항에 있어서,

상기 도메인 감성 사전을 구축하는 단계는,

상기 각 어휘에 대한 감성 점수를 시그모이드 함수를 사용하여 정규화하고, 상기 각 어휘에 대한 품사를 지정하고, 상기 지정된 각 어휘에 대한 품사에 따라 어휘별 감성 점수를 계산하는 단계

를 포함하는 감성 분석 방법.

【청구항 4】

제1항에 있어서,

상기 도메인 감성 사전을 구축하는 단계는,

상기 구축된 도메인 감성 사전의 점수와 범용 감성 사전의 점수를 비교하여 상기 구축된 도메인 감성 사전을 평가하는 단계

를 포함하는 감성 분석 방법.

【청구항 5】

제1항에 있어서,

상기 분류하는 단계는,

상기 구축된 도메인 감성 사전을 감성 스코어 임베딩 레이어로 변환하고, 상기 변환된 감성 스코어 임베딩 레이어와 워드 임베딩 레이어가 결합된 감성 분석 모델을 생성하고, 말뭉치에 대한 학습 데이터를 이용하여 상기 생성된 감성 분석 모델을 학습시키는 단계

를 포함하는 감성 분석 방법.

【청구항 6】

제5항에 있어서,

상기 분류하는 단계는,

상기 감성 스코어 임베딩 레이어를 통한 감성 스코어 임베딩을 통해 도메인 감성 사전을 벡터로 표현하고, 상기 워드 임베딩 레이어를 통한 워드 임베딩을 통해 각 어휘의 유사도를 반영하기 위한 벡터로 표현하는 단계

를 포함하는 감성 분석 방법.

【청구항 7】

제5항에 있어서,

상기 말뭉치에 대한 학습 데이터는,

상기 감성 분석 모델의 학습을 위해 데이터 전처리가 수행된 것을 특징으로 하는 감성 분석 방법.

【청구항 8】

제7항에 있어서,

상기 데이터 전처리는,

텍스트 데이터에서 특수 문자를 포함하는 불필요한 문자를 제거하는 데이터 정제 과정;

상기 불필요한 문자가 제거된 텍스트 데이터에서 불용어를 제거하는 불용어 제거 과정;

상기 불용어가 제거된 텍스트 데이터에서 형태소 정규화를 통해 어휘를 추출하는 어간 추출 과정;

상기 형태 정규화를 수행한 어휘마다 번호를 부여하여 토큰화된 어휘를 생성하는 토큰화 과정;

상기 토큰화된 어휘를 기반으로 감성 분석에 사용할 어휘를 토큰 번호로 변환하는 인코딩 과정; 및

상기 감성 분석에 사용할 어휘를 동일한 크기의 벡터로 변환하는 패딩 과정을 수행하는 것을 특징으로 하는 감성 분석 방법.

【청구항 9】

감성 분석 시스템에 있어서,

말뭉치를 기반으로 도메인 특성이 반영된 도메인 감성 사전을 구축하는 사전 구축부; 및

상기 구축된 도메인 감성 사전과 워드 임베딩이 결합된 감성 분석 모델을 이용하여 새로운 말뭉치로부터 도메인에 따른 감성을 분류하는 감성 분류부

를 포함하고,

상기 감성 분석 모델은, 말뭉치에 대한 학습 데이터를 이용하여 각 어휘의 도메인 감성 스코어와 각 어휘의 유사도가 반영되어 도메인에 따라 감성이 분류되도록 학습된 것을 특징으로 하는 감성 분석 시스템.

【요약서】

【요약】

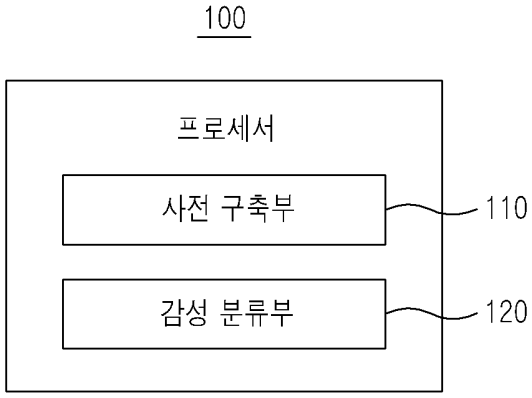
도메인 감성 사전과 워드 임베딩 기법을 결합한 감성 분석 방법 및 시스템이 개시된다. 일 실시예에 따른 감성 분석 시스템에 의해 수행되는 감성 분석 방법은, 말뭉치를 기반으로 도메인 특성이 반영된 도메인 감성 사전을 구축하는 단계; 및 상기 구축된 도메인 감성 사전과 워드 임베딩이 결합된 감성 분석 모델을 이용하여 새로운 말뭉치로부터 도메인에 따른 감성을 분류하는 단계를 포함하고, 상기 감성 분석 모델은, 말뭉치에 대한 학습 데이터를 이용하여 각 어휘의 도메인 감성 스코어와 각 어휘의 유사도가 반영되어 도메인에 따라 감성이 분류되도록 학습된 것일 수 있다.

【대표도】

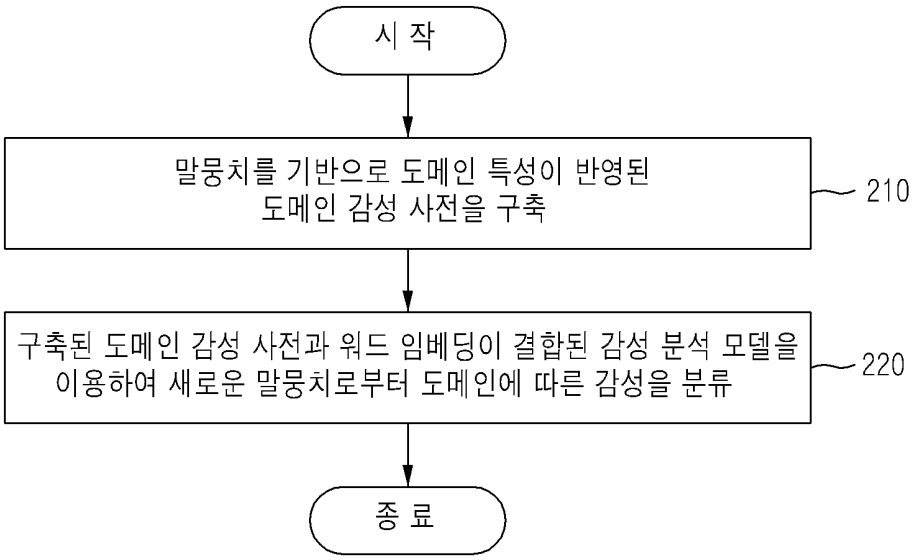
도 9

【도면】

【도 1】


100
프로세서
사전 구축부
110
감성 분류부
120

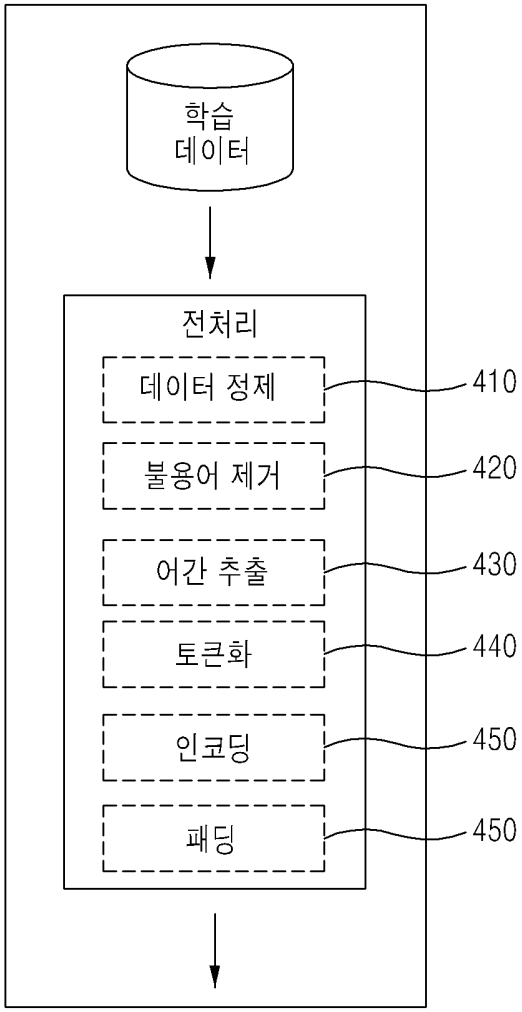
【도 2】


시 작
말뭉치를 기반으로 도메인 특성이 반영된
도메인 감성 사전을 구축
210
구축된 도메인 감성 사전과 워드 임베딩이 결합된 감성 분석 모델을
이용하여 새로운 말뭉치로부터 도메인에 따른 감성을 분류
220
종 료

【도 3】

	id	document	label	
0	9976970	아 더빙..진짜 짜증나네요 목소리	0	부정 (label = 0)
1	3819312	흠....포스터보고 초딩영화줄....오버연기조차 가볍지 않구나	1	긍정 (label = 1)
2	10265843	너무재밓었다그래서보는것을추천한다	0	부정 (label = 0)
3	9045019	교도소 이야기구먼..솔직히 재미는 없다..평점 조정	0	부정 (label = 0)
4	6483659	사이몬페그의 익살스런 연기가 돋보였던 영화!스파이더맨에서 늙어보이기만 했던 커스틴...	1	긍정 (label = 1)

【도 4】



【도 5】

Score (W_p)
L_{tp}
W_p
1 /2 = **0.5**
2
'재미있다'
Normalization
+
POS
tagging
[('아주', 'Noun'), ('**재미있다**', '**Adjective**'), ('책', 'Noun'),
('으로', 'Josa'), ('보다', 'Verb'), ('정말', 'Noun'),('**재미있다**', '**Adjective**')]
t
S_t
p
1
Adjective
아주 **재미있고** 책으로봐도 정말 **재미있다**

【도 6】

id	text	label	word	score
0	디 워랑 같은 부류 저예산 저렴한 CG 마구잡이 연출	0	저렴하다	-0.16667
1	저렴한 요괴 루협예로의 볼거리 그 이상의 철학적 내용을 담아내려 노력한 작품ㅋ	1	저렴하다	-0.06667
2	손 빈 이 천짓만 안 했어도 결말의 저렴함은 어찌할 수가 없다	0	저렴하다	-0.19167
3	나이가 들어도 여자는 여자다 시도는 좋으나 저렴한 표현 그녀의 소설과 난 상극인 듯	0	저렴하다	-0.28258
4	별점은 선택 안 함이 없네 그래도 다행인 건 집에 영화 다시 보기에 이 영화가 2천 원이라서 나름 저렴한 가격에 봤다는 거 안 다행인 건 그 돈마저 아깝다는 거	0	저렴하다	-0.354
5	ㅋㅋㅋ 미치겠네 ㅋㅋㅋ 지자 배우 연기력이 머리 정도까지 저렴한 영화가 있었나고 생각한다 사랑과 전쟁도 이것보단 연기 잘 할 듯 다 부숴버릴 거야 할 때 정말 모니터 부숴버릴 뻔했네 ;	0	저렴하다	-0.39567
6	너무 저렴하다 싸구려 쓰레기	0	저렴하다	-0.64567
7	저렴하게 표현된 성의 노예가 된 여자들	0	저렴하다	-0.78853
8	저렴한 작품	0	저렴하다	-1.28853
9	제작 비용 많이 들어 보이는데 저렴한 영화인 듯	0	저렴하다	-1.43139
10	저예산이라는 게 연기가 저렴해서 저예산이라는 건 아닐 텐데	0	저렴하다	-1.63139
11	감독이 천재구나 거기에다가 저렇게 저렴한 제작비로 만들 수도 있구나를 알게 한 영화	1	저렴하다	-1.54805
12	여자 출연자들 진짜 불쌍함 감독이 연기는 포기하고 저렴한 노출 가능 배우를 캐스팅한 느낌 대사를 평소 대화하는 것처럼 하면 되는데	0	저렴하다	-1.60068
13	저는 정말 재미있게 봤네요 Vdo 가격도 저렴해서 iptv로 구입해서 봤는 데야 하면서도 공포영화 보는듯하면서도 스릴도 있으면서 악간의 엘로에 두루두루 들어간 영화 같네요	1	저렴하다	-1.55306
14	저렴하게 만들어진 이상한 영화	0	저렴하다	-1.80306
15	저렴하게 포장해야 타란티노의 진가가 드러난다ㅋㅋ	1	저렴하다	-1.6364
16	뭔가 블록버스터의 스케일 영화 같긴 한데 저렴한 느낌이 들지	0	저렴하다	-1.7614
17	인내심 최대치로 참아가면서 끝까지 보면 결국은 욱 튀김 이런 저렴한 호러가	0	저렴하다	-1.87251
18	영화 참 저렴하네 완성도에서 너무 쳐진다	0	저렴하다	-2.03918
19	목소리 들으면 다 아는 것을 헛짓거리 하는 웃긴 영화 ㅋㅋ 거기다가 여자 주인공 졸라 걸레 무슨 남자를 바로 교체하는 저렴한 몸뚱어리를 보여주노 아님 여자란 동물이 원래 그런 건가 졸라	0	저렴하다	-2.06482
20	정수가 왜 이래 ㅋㅋ 나름 의미 있는 영화를 만들고 싶었나 본데 보면서 느낀 점이란 허술하고 저렴한 거지 같은 영화 ㅋㅋㅋ 마지막에 여자배우 중 손잡이에 손가락 안 넣고 있는 것도 압권	0	저렴하다	-2.10482
21	참 저렴한 이글 아이	0	저렴하다	-2.35482
22	평점 보고 깔짝 놀랐다 뭐냐 이 말도 안 되게 저렴한 평점은	1	저렴하다	-2.21196
23	대사가 왜 이렇게 저렴한 지참 보고 기본 나빠지는 영화임	0	저렴하다	-2.32307
24	저렴한 떡밥아 영화이지만 하우스 유직 좋고 딱히 볼 거도 없고 가끔 꼴리게도 해주	0	저렴하다	-2.4064
25	참 저렴하고 재미없는 코디니 영화	0	저렴하다	-2.6064
26	허술하고 놀어지고 식상하고 저렴하다	0	저렴하다	-2.8564
27	이걸 두 번 보다니 아까운 내 시간 아직도 결말과 이 작가와 감독의 의도를 모르겠네 중 등장인물 16 17명 중 맞으면 임무 끝 주문 외우면 임무 끝 제작비 저렴하겠네 세상에서 출연배우 돈	0	저렴하다	-2.90186
28	참 저렴한 영화 시끄러운 스페인어에다가 이유도 모르고 갇히는 상황 셀 수 없이 흔들리는 카메라 등 전반적으로 영화라고 부르기에도 참 뭐 한 영상물이다 중간에 끊고 싶었지만 마지막까	0	저렴하다	-2.93519
29	액션도 없고 판타지도 없고 이야기도 없는 저렴한 뮐름 영상	0	저렴하다	-3.0463
30	글발이 너무 저렴해 보이고 기대했더니만 하이틴 영화였네	0	저렴하다	-3.1713
31	주인공들의 연기력 뿐 아니라 영상미도 대단하고 ost도 참 차분하고 무게감 있어서 19금 영화지만 저렴한 느낌은 전혀 없었음	1	저렴하다	-3.13559
32	재밌는데요 프랑스 뮤지컬 공연을 극장에서 저렴하게 뮤지컬 보는 거에 비해 저렴하는 말임 볼 수 있으니 좋아요	1	저렴하다	-3.05867
33	재밌는데요 프랑스 뮤지컬 공연을 극장에서 저렴하게 뮤지컬 보는 거에 비해 저렴하는 말임 볼 수 있으니 좋아요	1	저렴하다	-2.98174

【도 7】

Vocabulary	Score
훌륭하다	205.6197
고결하다	1.1428
살짝살짝	0.1166
허우적허우적	-0.5000
교활하다	-2.0000
허접스럽다	-4.4000

Normalization

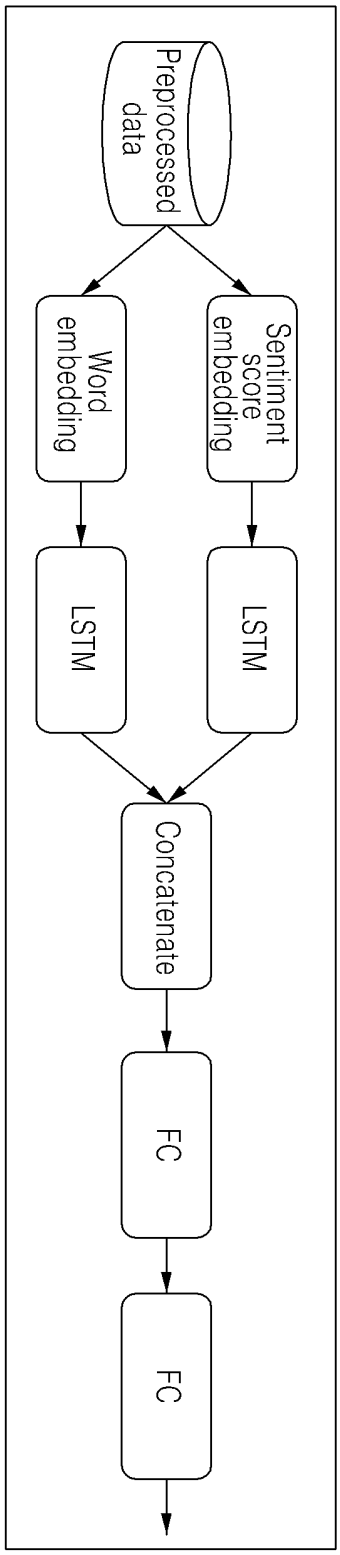
Vocabulary	Score
훌륭하다	0.9834
고결하다	0.7582
살짝살짝	0.5291
허우적허우적	0.3775
교활하다	0.1192
허접스럽다	0.0121

【표 8】

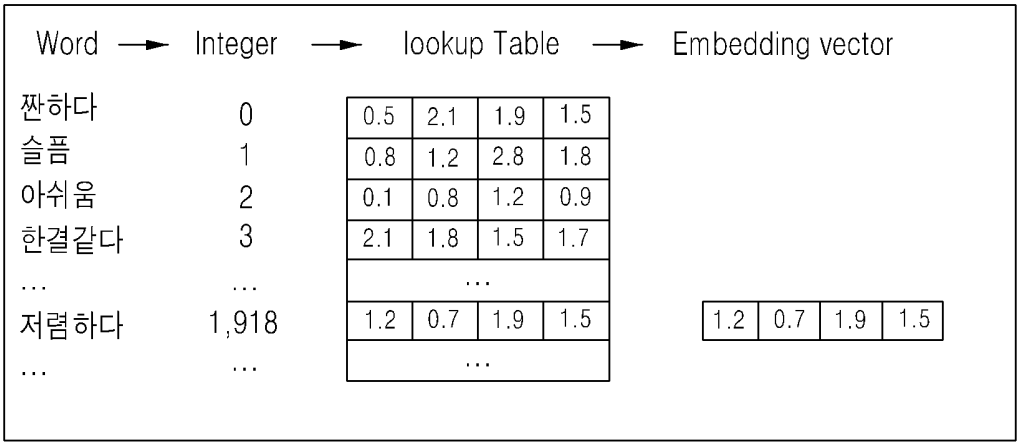
Vocabulary	Score	
	Proposed	KNU
짠하다	0.999	-2
슬픔	0.999	-2
아쉬움	0.967	-1
한결같다	0.295	1
저렴하다	0.048	1

Vocabulary	Score	
	Proposed	KNU
죽이다	0.914	-1
피로	0.886	-1
통증	0.745	-1
전문가	0.112	1
개선	0.001	2

【도 9】


Preprocessed data
Word embedding
Sentiment score embedding
LSTM
LSTM
Concatenate
FC
FC
900

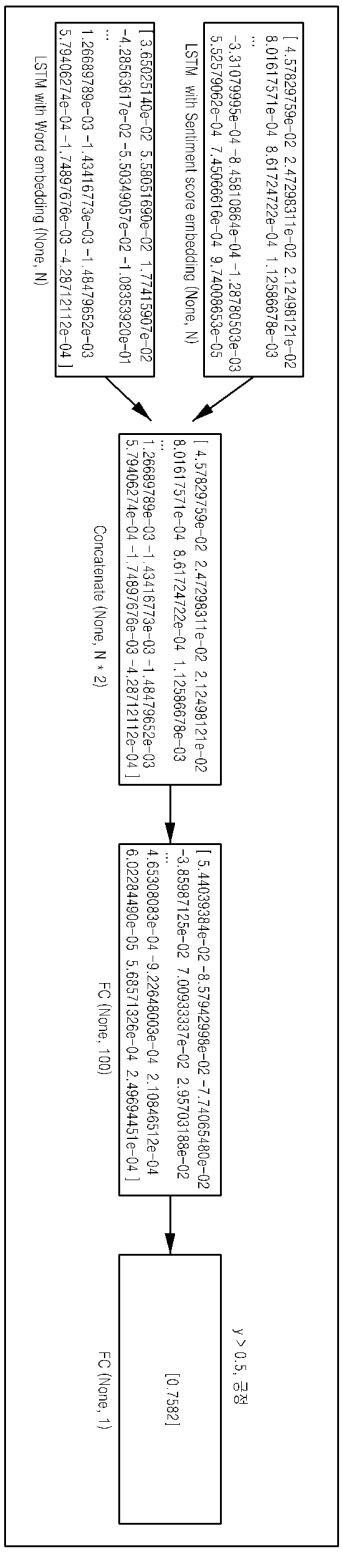
【도 10】


Word
Integer
lookup Table
Embedding vector
짠하다
슬픔
아쉬움
한결같다
…
저렴하다
…
0
1
2
3
…
1,918
…
0.5 2.1 1.9 1.5
0.8 1.2 2.8 1.8
0.1 0.8 1.2 0.9
2.1 1.8 1.5 1.7
…
1.2 0.7 1.9 1.5
…
1.2 0.7 1.9 1.5

【도 11】

word	Integer (Encoding)	lookup table (Embedding vector)
보고	1	1
역기	2	1
보다	3	1
추천	4	1
이쁘다	5	1
	⋮	
전술	16315	0.498567809
말다툼	16316	0.498553245
가위바위보	16317	0.498543053
블로거	16318	0.49851436
내려다보다	16319	0.498514436
	⋮	
별로	33953	7.44E-172
지루하다	33954	4.99E-178
쓰레기	33955	8.90E-238
아깝다	33956	1.58E-281

【도 12】


[4.57829759e-02 2.47298311e-02 2.12498121e-02
8.01617571e-04 8.61724722e-04 1.12586678e-03
...
-3.31079995e-04 -8.45810684e-04 -1.28780503e-03
5.52579062e-04 7.45066616e-04 9.74008653e-05
LSTM with Sentiment score embedding (None, N)
[3.65025140e-02 5.58051690e-02 1.77415907e-02
-4.28563617e-02 -5.50349057e-02 -1.08353920e-01
...
1.26689789e-03 -1.43416773e-03 -1.48479652e-03
5.79406274e-04 -1.74897676e-03 -4.28712112e-04]
LSTM with Word embedding (None, N)
[4.57829759e-02 2.47298311e-02 2.12498121e-02
8.01617571e-04 8.61724722e-04 1.12586678e-03
...
1.26689789e-03 -1.43416773e-03 -1.48479652e-03
5.79406274e-04 -1.74897676e-03 -4.28712112e-04]
Concatenate (None, N * 2)
[5.44039384e-02 -8.57942998e-02 -7.74065480e-02
-3.85987125e-02 7.00933337e-02 2.95703188e-02
...
4.65308088e-04 -9.22648003e-04 2.10846512e-04
6.02284490e-05 5.68571326e-04 2.49694451e-04]
FC (None, 100)
[0.7582]
FC (None, 1)
y > 0.5, 긍정