8-K 공시 문서의 분산 표상 기반 감성 분석을 통한 주가 방향 예측

Sentiment Analysis of Companies' 8-K Financial Reports based on Distributed Representation for Stock Price Change Prediction

김미숙 박은정 조성준

mskim@dm.snu.ac.kr epark@dm.snu.ac.kr zoon@snu.ac.kr

December 28, 2015

Abstract

금융 시장에서 주가 예측을 하기 위해 다양한 연구가 이루어지고 있다. 과거에는 계량적인 지표를 이용한 연구가 주를 이루었으나, 현재는 텍스트 정보까지 활용하여 다양한 알고리즘을 적용한 연구가 이루어지고 있다. 주가를 예측하는 모델이 현업의 실무자 및 경영진들에게 주어질 수 있다면 이러한 모델을 기반으로 의사 결정을 할 수 있게 되고, 좀 더 객관적이고 정확한 판단을 할 수 있는 기반이 될 수 있다. 그러나 기존의 연구들은 도메인에 관계 없이 대량의 텍스트 문서를 활용하여 주가 방향을 분류하는 문제를 해결해왔으나, 단어가 도메인에 따라서 다르게 해석 될 수 있기 때문에 여러 도메인이 섞여 있는 대량의 문서를 한번에 학습하여 특정 도메인에 적용하고자 하면 오히려 성능이 저하될 수 있다. 따라서 본 연구에서는 8-K 공시 문서를 바탕으로 특정 도메인에 속하는 기업의 공시 문서만 샘플링하여 주가의 UP, DOWN을 예측하는 문제를 해결해보고자 한다. 분산 표상(Distributed Representation) 기반으로 감성 분석(Sentiment analysis)을 하여 모델 기반 및 시각화 기반의 주가 방향 예측을 한다. 모델 기반 주가 예측을 위해 총 6가지 분류 알고리즘을 사용하여 분류 성능을 비교했고, 분산 표상 방법으로 가장 높은 분류 성능을 얻을 수 있었다. 시각화 기반 주가 예측은 2008 ~ 2009년도의 금융 위기 때의 특정 기업의 공시 문서를 2달 간격으로 나누어 기간에 따라 공시 문서의 감성을 시각화 하였고, 이를 실제 주가 흐름과 비교했을 때 주가 흐름과 공시 문서의 감성이 비슷한 경향을 띄는 것을 확인했다. 이와 같이 모델 기반으로 주가 예측을 하게 되면 예측 정확도가 정량적으로 구해지면서 현업의 실무자들에게 정보를 제공해 줄 수 있을 것으로 기대할 수 있고, 반면 시각화 기반으로 주가 예측을 하게 되면 시각화를 통해 주가의 흐름을 한 눈에 보여 줄 수 있어 경영자들에게 의사 결정 지원이 가능할 것으로 예상된다. 이처럼 분산 표상 기법을 활용하여 정량적 및 정성적인 주가 방향 예측을 통해 금융 시장의 다양한 통찰(insight)를 얻을 수 있을 것으로 기대한다.

키워드: stock price change prediction, distributed representation, sentiment analysis, 8-K financial reports, model-based, visualization-based

1 Introduction

금융 분야에서의 주가를 예측하고자 하는 노력은 과거부터 현재까지 끊임없이 연구되어 왔다. 과거 연구에서는 주가 예측을 위해서 계량적 지표를 이용한 수학적 모델을 이용했는데, 이 모델은 주가의 과거 시계열 정보, 미시 경제 지표 및 거시 경제 지표 등의 변수를 사용했다[8]. 효율적 시장 가설(Efficient Market Hypothesis)에서는 시장을 완벽하게 예측하는 것이 불가능하다고 주장하지만, Fama(1968)[7]는 어느 정도의 예측은 가능하다고 주장하고 있다. 주식 가격 예측은 다양한 접근으로 이루어져 왔는데, 대표적인 접근 방법으로 과거 주식 가격에 autoregressive integrated moving average(ARIMA)와 같은 시계열 모델을 적용하여 주식 가격 예측(Stock price prediction) 연구가 이루어져 왔다[2]. 또한 인공지능 분야의 알고리즘, 예를 들면 신경회로망(neural network), 의사결정나무(decision tree), 서포트 벡터 머신(support vector machine) 등을 적용하여 주가 예측 정확도를 향상시켜왔고, 이러한 시도들은 주가 예측에 눈에 띄는 성과를 나타냈다[6, 1].

이러한 계량적인 변수들을 이용하여 다양한 수학적 모델 및 인공지능 모델이 적용되어 왔지만, 제한된 변수들만을 이용해서는 더 이상의 예측 성능 향상을 기대하기 어려워졌다. 그러나 이러한 모델들은 실제 매매에 있어서 한계가 있어 현업에서 활용되지 않았다. 따라서 현재 연구들에서는 기존의 계량적 변수뿐만 아니라 뉴스, SNS, 공시 등의 텍스트 데이터를 분석하여 주가 예측 연구가 이루어지고 있고, 주가 예측 성능이 향상되어 오고 있다. [10]. 텍스트 데이터에서 나타나는 특정 단어나, 문장, 문단, 문서 등의 범주는 각각 투자자 등의 감성(sentiment)을 내포하고 있고, 이러한 감성 분석은 금융 시장에 있어서 주가 예측의 중요한 방법으로 여겨지고 있다. 또한 주가 예측 성능의 향상을 위해 복잡한 모델, 예를 들면 딥 러닝(deep learning) 등을 이용한 모델들도 최근에 다양하게 제안되고 있다[15]. 그러나 하나의 문서에 사용되는 단어가 도메인에 따라서 다르게 해석될 수 있어, 여러 도메인이 섞여 있는 문서를 한번에 학습하게 될 경우 오히려 성능이 떨어질 수 있다. 따라서 특정 도메인을 구분하여 모델을 학습하고 이를 바탕으로 성능을 분석하는 연구가 이루어져야 한다.

본 연구에서는 미국 S&P 1,500 기업의 8-K 공시 문서를 이용하여 공시 이후 주가의 UP, DOWN을 다음과 같은 과정을 통해 예측하고 추가적인 통찰(insight)을 얻고자 한다. 8-K 공시 문서를 분산 표상(Distributed Representation) 기반으로 감성 분석(Sentiment Analysis)을 하여 모델 및 시각화를 통한 주가 예측에 사용하고자 한다. 먼저 모델 기반의 주가 예측은 다양한 방법론을 통해 정량적인 예측력을 제시하고 실제 현업 실무자들에게 의사결정 기반을 제공하고자 한다. 유니그램(unigram)을 통한 단어를 입력 변수로 사용한 모델을 바탕으로 주가의 방향을 예측하여 이를 베이스 라인(Baseline)으로 사용하고, Multinomial Naïve Bayes(MNB), Support Vector Machines(SVM) 및 Wang([16])의 NBSVM(SVM with NB feature)을 공시 문서에 적용하여 그 결과를 비교해보고자 한다. 또한 분산 표상 방법 자체가 갖고 있는 예측 성능을 이용하여 주가 예측을 해보고 다양한 방법론의 성능을 비교 분석하고자 한다. 다음으로 시각화 기반의 정성적인 주가 예측을 통해 경영자들의 의사 결정에 도움이 되는 분석을 하고자 한다. 분산 표상 방법을 통해 문서와 감성 클래스를 동일 공간에 표현할 수 있고, 이 문서들을 공간에 시각화하여 금융 시장의 통찰을 얻고자 한다.

본 연구에서 다룰 내용은 다음과 같다. 2절 에서는 텍스트 데이터를 활용한 주가 예측에 대한

연구 및 텍스트 데이터의 분석 방법론에 대한 기존 연구들에 대해 살펴보고자 한다. 3절에서는 본 연구에서 사용하는 방법론에 대해 소개하고, 4절에서는 사용하는 데이터에 대한 description과 예측 정확도 및 시각화 결과를 논의할 것이다. 끝으로 본 연구에 대한 결론 및 제언을 5절에서 기술할 것이다.

2 Related Work

금융 시장에서 계량적 데이터뿐만 아니라 텍스트 데이터를 이용하여 주가 예측에 대한 연구가 활발하게 이루어지고 있다. 전체 소비자의 감성을 나타내어 주는 뉴스뿐만 아니라 개별적인 감성이 드러나있는 SNS, 블로그 등의 데이터까지 이용하여 주가 예측에 활용되고 있다[3]. 더불어 텍스트 데이터의 분석 방법론이 다양하게 개발 및 적용되고 있다[15, 12].

텍스트 데이터를 이용하여 주가 예측에 활용되기 시작할 때 다양한 텍스트 데이터, 예를 들면 뉴스, 중요 대화록, 공식 문서뿐만 아니라 블로그, SNS 등 개인적인 텍스트 문서까지 주가 예측에 활용되었고, 많은 연구에서 성능 향상을 이루었다. Lee(2014)[10]는 미국의 S&P 1500 기업에 대한 공시 문서를 이용하여 Earning surprise, Recent movement, Volatility, Event category 등 계량적인 지표들만 사용한 기존 분석 보다 언어적 요인(Linguistics features)을 사용할 때 주가 예측 성능이 향상되는 것을 실험을 통해 확인했다. 언어적 요인을 모델에 추가하여 Random Forest(RF)를 통해 3개의 클래스(UP, STAY, DOWN)로 이루어진 데이터를 분류했고 그 성능을 확인했다. 그 결과 계량적인 지표들만 이용했을 때는 50.1%의 정확도를 얻을 수 있었고, 언어적 요인을 추가했을 때에는 55.5%까지 그 성능이 향상되었다.

Bollen(2011)[3]은 twitter 데이터를 2가지 방법으로 public mood를 측정했다. 먼저 Opinion-Finder를 통해 public mood를 2개의 차원(positive vs. negative)로 표현했고, 둘째로 google-profile of Mood states(GPOMS)를 이용하여 6개의 차원 (calm, alert, sure, vital, kind, and happy)로 public mood를 측정했다. 이 변수를 이용하여 Granger causality analysis와 Self-Organizing Fuzzy Neural Network 방법론을 통해 DJIA의 종가의 UP, DOWN 예측에 사용했고, 그 결과 86.7%의 예측 성능을 얻을 수 있었고, Mean Average Percentage Error(MAPE)가 6%이상 감소하는 결과를 나타냈다.

Liu(2015)[11]는 투자자의 감성과 주식 시장 유동성(stock market liquidity)의 time-series variation 사이의 관계에 대해 연구했고, Granger-causes test를 통해 그 영향이 유의함을 검증했다. Time series regression을 통해 투자자의 감성이 특히 주식 시장의 거래량 증가와 price effect의 감소에 영향을 주는 것을 확인했다. 또한 투자자의 강한 감성은 간접적으로 영향을 주기도 하지만 직접적으로 영향을 주는 것을 데이터를 통해 검증 했다.

금융 시장에서 텍스트 데이터를 이용한 주가 예측은 계량적 지표만을 사용했을 때 보다 훨씬 좋은 성능이 나타나는 것이 여러 연구를 통해 이루어짐에 따라 이 성능을 향상시키기 위한 다양한 알고리즘들이 개발 및 적용되고 있다. Mass(2011)[12]는 단어 표현(word representation) 방법을 이용하여 감성의 극성(polarity)을 예측했다. 비슷한 감성을 갖는 단어들은, 단어가 공간상으로 벡

터로 표현될 때 비슷한 위치에 표현되는 특징을 이용하여, 의미(semantic) 정보뿐만 아니라 감성 정보를 감성의 극성 예측 모델에 입력 변수로 사용하여 분석했다. 영화 리뷰 데이터로 많이 사용되는 IMDB 데이터를 이용하여 5,000개의 단어 사전을 구축하여 사용했다. 로그 선형(log-linear) 모델을 활용하여 감성의 극성을 예측하여 LDA(Latent Dirichlet Allocation)보다 향상된 성능을 얻을 수 있었고, 감성 정보를 활용함으로써 단어의 의미를 예측할 수 있는 단어들뿐만 아니라 문장에서 비슷한 단어를 대체할 수 있는 단어까지 비슷한 단어로 공간상에 전사되는 것을 확인 할 수 있었다. 간단한 감성 정보의 활용만으로 성능향상을 이루어낸 점에서 향후 활용가능성이 크다는 의의가 있다.

Wang(2012)[16]는 일반적으로 텍스트를 분류하는 문제에 있어서 베이스 라인으로 활용되는 Naïve Bayes(NB), Support Vector Machines(SVM)이 어떤 데이터 셋이나 조건에서는 state-of-the-art보다 성능이 좋게 나타나는 것을 바탕으로 복잡한 모델이 항상 좋은 성능을 나타내는 것이 아니라고 주장했다. 사용한 데이터는 데이터 분석 시 많이 사용되는 RT-s, CR, MPQA, IMDB 등 다양한 데이터 셋을 구축했고, 각 문서 길이가 긴 문서와 짧은 문서로 구분하여 실험했다. 비교한 방법론들은 최근에 많이 개발되고 있는 복잡하고 rule-based 모델과 같은 Tree-CRF[14], Recursive Autoencoder (RAE) [15], RAE-pretrain[5], voting, rule 등의 방법론과 비교했다. 그 결과 NB의 요소를 이용한 SVM의 성능이 대체로 좋게 나타나는 것을 알 수 있었다. NB의 경우는 대체로 문서 길이가 짧은 문서에서 성능이 좋고, SVM의 경우에는 문서의 길이가 긴 문서에 대해서 성능이 좋다.

Socher(2013)[15]은 단어 하나를 벡터 공간에 표현하는 것은 단어 사이의 관계를 표현하지 못하는 한계점을 극복하기 위해 단어 간의 관계를 감성 트리뱅크(treebank)를 통해서 표현하여 분석하는 방법을 제안했다. Recursive Neural Tensor Network(RNTN)를 제안하여 높은 노드에 있는 단어의 벡터를 그 아래의 노드들을 이용하여 계산하여 하나의 단어 벡터가 주변 단어를 고려하여 표현될 수 있도록 알고리즘을 개발했다. 그 결과 80.7%의 정확도를 얻을 수 있었고, 특히 부정문이 갖는 까다로운 감성의 극성을 다른 알고리즘에 비해 정확하게 분류할 수 있었다.

3 Methods

본 연구에서는 서론에서 언급한 바와 같이 주식 가격 방향 예측 문제를 두 가지 방법으로 해결해보고자 하며, 그 과정을 그림 1에 나타내었다.

그림 1과 같이 분산 표상을 통해서 감성 분석을 수행하고, 이 감성 분석 결과를 이용하여 모델 기반 및 시각화 기반으로 주식 가격 방향 예측을 할 것이다. 모델 기반을 이용하여 예측 정확도를 구하여 정량적인 예측력을 보여주고, 시각화 기반을 이용하여 정성적으로 주식 가격의 변화를 보여주고자 한다.3.1절에서는 감성 분석을 하기 위해 사용하는 분산 표상 방법론에 대해서 설명하고, 3.2절에서는 시각화를 위해 사용하는 틀에 대한 예시와 설명을 할 것이다. 그리고 3.3절에서는 모델 기반을 위해서 사용하는 방법론들에 대한 설명과 장단점을 다룰 것이다.

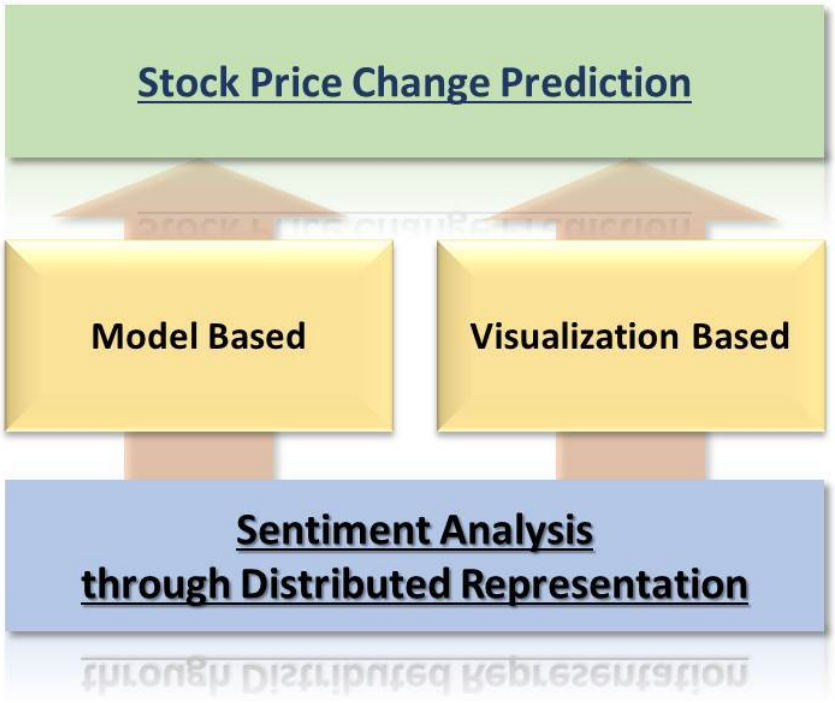


그림 1: Diagram of Stock Price Change Prediction

3.1 Distributed Representation

문서를 분류하는 문제는 텍스트 마이닝 연구에서 매우 중요한 요소이다. 따라서 분류 문제를 위해 문서를 어떻게 표현할 것인지 적절한 표현 방법을 찾는 것 역시 중요한 연구 분야이다. 문서를 표현하는 방법으로 bag-of-words와 같은 방법이 제안되었는데, 이 방법은 단어의 수가 증가함에 따라 차원이 매우 커지게 되고, sparsity 문제가 생기게 된다. 따라서 문서를 차원을 작고 dense하게 표현하는 방법들이 제안되고 있다. 이런 방법을 분산 표상이라고 하는데, 이는 discrete objects까지도 연속 공간으로 표현이 가능하다. 각 데이터를 연속 공간으로 표현하게 되면 데이터 사이의 유사도를 계산할 수 있어 문서간 유사도를 계산하는 등 다양하게 활용될 수 있다. 또한 데이터가 많지 않을 때에도 어느 정도 학습 성능이 보장된다. 이러한 장점을 바탕으로 여러 분산 표상 방법들이 연구되고 있고 대표적인 모델이 word2vec이다[13]. Word2vec이란 단어를 벡터 공간으로 임베딩(embedding) 시켜서 표현하는 방법으로 두 개의 층을 갖는 신경회로망의 형태의 모델이다. Word2vec은 크게 두 가지 방법으로 나누어 지는데 입력 텍스트가 들어오게 되면 중심 단어를 바탕으로 주변 단어를 예측하거나 주변 단어들을 바탕으로 중심 단어를 예측하는 방법이 있다. 전자를 Skip-gram, 후자를 CBOW이라고 한다. 이 중 Skipgram에 대해서 다음 문단에서 자세하게 설명했다.

Skipgram은 유사도를 정의하기 어려운 discrete 데이터 간의 거리를 학습하기에 적합한 딥러닝(deep learning) 방법이다. Skipgram은 discrete 자료를 연속형 벡터로 표현하는데, 비슷한 문맥의 단어들은 연속형 벡터 공간상에서 가까이 위치하도록 학습한다. 간단한 skipgram을 통해서 문장의 단어들이 공간으로 전사되는 모습을 그림 2에 나타내었다.

skipgram은 주어진 단어를 이용하여 해당 단어 주위에 등장하는 단어를 예측하는 encoder와

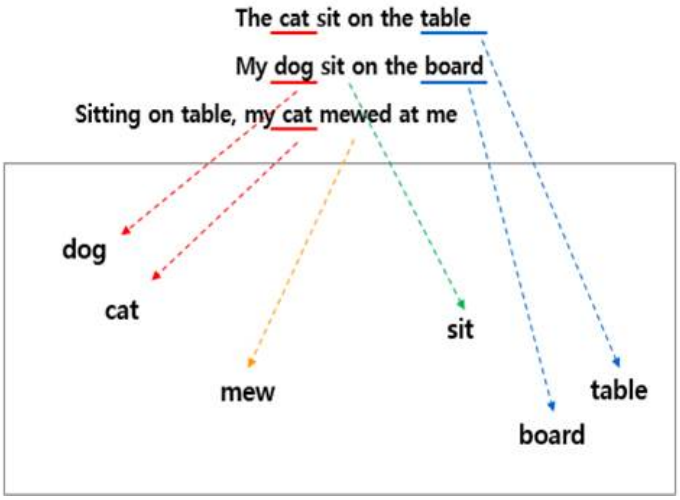
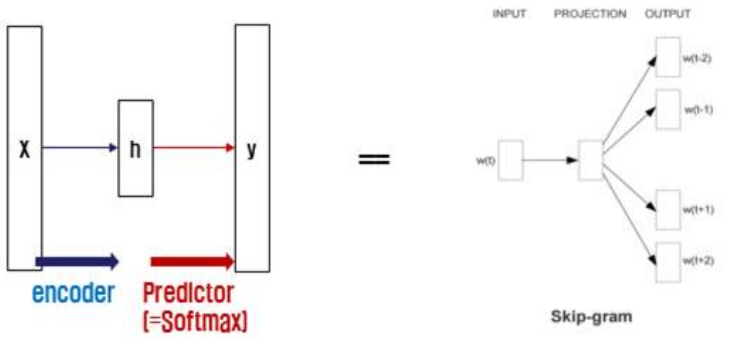


그림 2: Skipgram을 통한 단어의 공간 전사

predictor로 구성된 신경회로망 모델을 학습시킨다. Encoder는 discrete하게 표현된 입력 단어에 대하여 연속형 벡터로 변환하고, 이 변환된 벡터는 predictor를 통해 주위의 단어들을 예측한다. 주위의 단어 분포가 비슷한 두 단어는 서로 비슷한 연속형 벡터를 지니도록 encoder가 학습되면서 단어들 간의 거리가 정의된다. 이 과정을 그림 3와 같이 표현했다.



$x \in R^{V\times 1}$ One hot representation word vector. Input layer size equals with no. of terms

$h \in R^{d\times 1}$ Encoded word vector

$y \in R^{V\times 1}$ Context prediction prob. Output layer size also equals with no. of terms

그림 3: Skipgram 모델

매 시점마다의 데이터 값을 단어로 생각한다면, 범주형 데이터로 모든 변환된 자동차 주행 데이터도 단어들의 집합으로 생각할 수 있다. Skipgram을 적용한다면 유사한 주행 상황을 연속형 공간 벡터를 통해서 정의 할 수 있게 된다. 이를 통해 주행 데이터를 연속형 공간 벡터들의 시계열 데이터로 표현할 수 있게 된다.

이러한 word2vec에 paragraph vectors(PV) 정보를 그림 4와 같이 입력층에서 사용하게 되면, 단어와 문서 벡터를 동시에 학습하여 같은 공간에 표현 할 수 있는 장점이 있다[9].

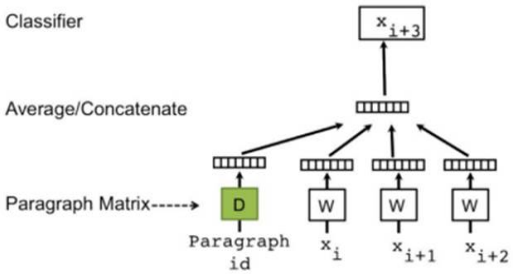


그림 4: Distributed model using paragraph vectors

그러나 word2vec이나 PV의 경우에는 분류 목적에 관계없이 같은 형태로 표현되는 데, 종종 분류 목적에 따라 다른 형태로 표현되는 것이 요구되는 경우도 존재한다. 예를 들면, word2vec은 의미와 구문 형태가 비슷한 단어를 벡터 공간에 가깝게 위치하게 하지만, 감성 분석 문제의 경우에는 "good"과 "bad"가 분리가 되어 벡터 공간상에 다른 곳에 위치하는 형태로 표현되기를 원한다. 따라서 이러한 문제를 해결하기 위해서 클래스 정보를 같이 학습하는 supervised PV 방법을 사용한다. 즉, supervised PV는 단어와 문서 벡터뿐만 아니라 클래스 벡터까지 입력층에서 사용하여 학습하게 되면 분류 목적이 달라지더라도 다른 형태의 결과로 표현될 수 있다. 따라서 본 연구에서는 이 방법을 이용하여 문서의 감성 분석을 하고 분류 및 시각화를 수행할 것이다.

3.2 Visualization

본 연구에서는 분산 표상을 통한 감성 분석 결과를 시각화로 보여주고자 한다. 텍스트 문서들을 분산 표상으로 감성 분석을 하게 되면 텍스트 문서 안의 단어, 문장, 문서, 클래스 정보까지 모두 같은 차원의 벡터로 표현할 수 있게 된다. 이때 보통 사용하는 차원이 100차원 이상인 경우가 많기 때문에 이를 바로 시각화 할 수 없다. 따라서 차원 축소 기법을 이용해서 시각화를 해야 한다. 본 연구에서는 차원 축소 기법 중 주성분 분석 기법을 이용하여 두 개의 주성분 변수를 찾고, 이 두 개의 변수를 이용하여 2차원 공간에 표현하고자 한다. 주성분 기법이란 모든 변수들의 변동(공분산, 상관계수)을 이용하여 변동을 가장 잘 설명하는 주성분 변수를 차례로 찾는 방법으로 주성분 변수는 기존의 변수들의 조합으로 나타내어진다. 예를 들어 각 변수들이 다음 그림 5와 같이 a, b의 축을 기준으로 나타내어 진다면, 주성분 분석 기법을 통해 중심축을 a', b'으로 이동시켜 재배치할 수 있다.

이와 같이 주성분 분석으로 고차원 벡터를 저차원 벡터로 나타내어 시각화 할 수 있다. 클래스 정보, 단어, 문서를 하나의 공간에 나타낸 그래프의 예시를 그림 6에 나타내었다.

위의 그림에서 x축에 'c1', 'c-1'는 클래스 정보를 나타낸 것으로 시각화가 잘 보이게 하기 위해 클래스 정보를 x축으로 고정시켜 단어와 문서를 벡터 공간에 표현했다. 파란색으로 표현된 긴 텍스트는 문서의 감성을 분석하고, 이 문서를 벡터로 표현하여 공간에 표현한 것이고, 녹색으로 표현된 단어들은 단어의 감성 결과를 벡터 공간에 나타낸 것이다. 위의 그림은 문서의 다수가 중립적인 문서인 것을 알 수 있고, 단어 중 china, korea, sinking 등이 샘플 도메인에서는 부정적인 단어로

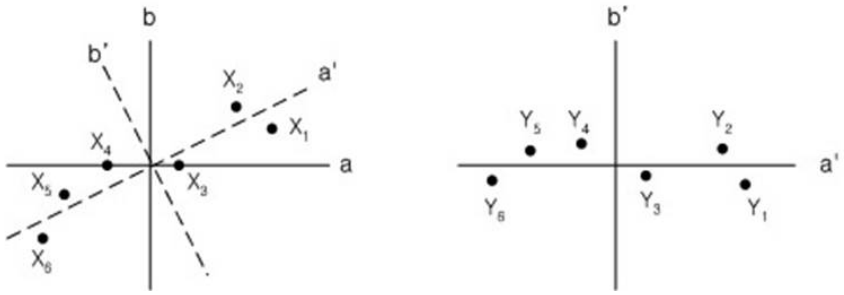


그림 5: 주성분 분석

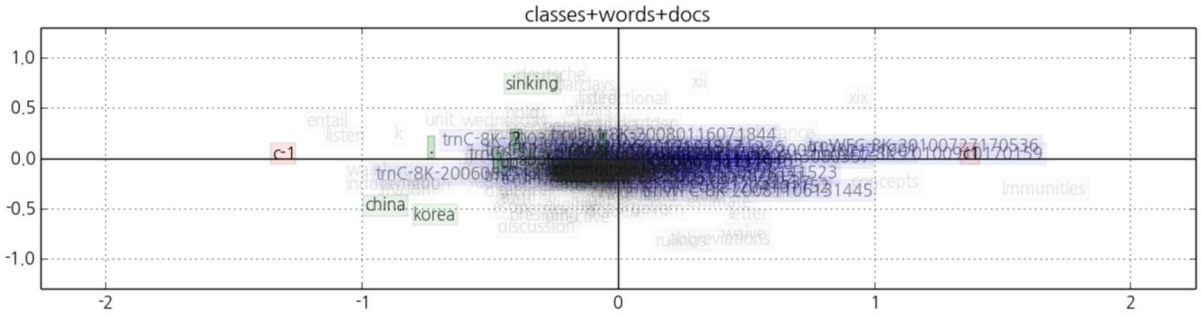


그림 6: Example of Visualization

나타나는 것을 알 수 있다. 이와 같이 단어, 문서, 클래스 정보를 하나의 공간에 나타내어 문서가 어떤 감성을 나타내는지를 위의 툴을 바탕으로 시각화하여 표현할 것이다.

3.3 Model Based Prediction

모델 기반의 주가 예측을 위해서 다양한 방법론을 사용하여 예측력을 높이고자 한다. 최근 텍스트 분류 문제에서 많이 사용되고 있는 알고리즘들을 사용하여 정량적인 예측 정확도를 제공하고, 본 연구에서 사용할 알고리즘들은 다음과 같다. 먼저, 유니그램을 이용하여 파싱(parsing)한 단어를 입력 변수로 사용하여 분류 모델에 적용한다. 이때 사용하는 분류 모델은 간단하고 계산 속도가 빠른 로지스틱 회귀모델(LR)과 높은 예측 성능을 갖고 있는 RF 모델을 통해 주가의 방향을 예측하고 이를 베이스 라인으로 사용한다. 또한 Multinomial Naïve Bayes(MNB), Support Vector Machines(SVM) 및 Wang(2012)[16]의 NBSVM(SVM with NB feature)을 공시 문서에 적용하고, 마지막으로 3.1절에서 제시한 분산 표상의 분류 성능까지 그 결과들을 비교해보고자 한다. 각 알고리즘의 자세한 설명은 3.3.1 ∼ 3.3.5절에서 다룰 것이다.

3.3.1 Random Forest

RF는 앙상블(ensemble) 방법 중 하나로써, 여러 개의 의사결정나무를 앙상블하여 하나의 분류 모델을 만드는 방법이다. 학습 데이터 셋은 $x_i, y_i, i = 1, \ldots, N$와 같이 구성되고, 이 데이터 셋을 이용하여 B개의 트리를 만든다. 이때 각 트리는 중복 추출(replacement)를 가능하게 하여 B개의 random sample set에 대해서 B개의 트리를 만든다. 학습이 끝난 뒤에, 모든 의사결정나무에서 나온

결과를 투표하고 가장 많이 등장한 클래스로 할당하여 분류한다[4]. RF는 분류 모델 중에서 매우 정확한 분류 성능을 내는 알고리즘 중 하나이다. RF는 높은 정확도뿐만 아니라 효율적으로 학습되며, 분류에 가장 중요한 변수들에 대한 정보도 추정 할 수 있다는 장점이 있다. 반면, 이상치를 포함하고 있는 데이터 셋에 대해서는 분류할 때 오버핏팅(overfitting)이 발생할 수 있다는 단점이 있다.

3.3.2 Logistic Regression

LR은 출력 변수가 범주형 변수 일 때 많이 사용되는 회귀 모형으로 로지스틱 함수(logistic function)를 통해 발생 확률을 예측할 수 있다. 즉, 어떤 사건이 발생여부를 직접 예측하는 것이 아니라, 실제 발생할 확률 예측하는 모델이다(Cox, 1958). X=x에서 Y=0, Y=1일 확률의 비율인 Odds ratio의 로그 값을 선형 회귀 모형으로 예측하는데, 이는 식 (1)과 같다.

$$y_i = log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \epsilon_i, \epsilon_i \sim iidN(0, \sigma^2) \tag{1}$$

위의 회귀식을 p_i에 관한 식으로 나타내어 이를 로지스틱 함수라고 정의할 수 있고, 식 (2)과 같이 나타낼 수 있다.

$$\hat{p} = \hat{P}(Y=1|x) = \frac{exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_n x_n)}{1 + exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_n x_n)} \tag{2}$$

위의 확률 값은 $0 \sim 1$의 값을 갖는데, 기준값(criteria) α를 정하고 $p \geq \alpha$이면 클래스 1, $p < \alpha$이면 클래스 0로 분류한다. LR는 계산 비용이 적고, 구현하기가 쉽고, 어떤 변수가 영향을 많이 주는지 등의 결과 해석이 쉬운 장점이 있지만, 오버핏팅 경향성을 가지고 있어 정확도가 낮게 나올 수 있는 단점이 있다.

3.3.3 Multinomial Naïve Bayes

MNB는 데이터가 주어졌을 때 각 변수들이 독립이라는 가정하에서 어떤 클래스에 속할 확률을 나타내는 조건부 확률을 이용한 분류 모델이다. n개의 변수로 구성된 데이터가 있고, K개의 클래스가 있다고 했을 때, 클래스 k에 속할 조건부 확률을 식(3)과 같이 나타낼 수 있다.

$$p(C_k|x_1, x_2, ..., x_n) = p(C_k|x) = \frac{p(C_k)p(x|C_k)}{p(x)} \tag{3}$$

위의 식에서 분모는 관찰된 데이터에서 추출하는 값이므로 클래스에 관계없기 때문에, 분자 부분을 이용하여 확률 모델을 만들 수 있고, 각 변수들이 독립이기 때문에 그 형태는 식(4)와 같이 표현할 수 있다.

$$p(C_k, x_1, x_2, ..., x_n) = p(C_k)p(x_1|C_k)p(x_2|C_k)...p(x_n|C_k) = p(C_k)\Pi_{i=1}^{n} p(x_i|C_k) \tag{4}$$

식(3)과 식(4)의 값은 비례하기 때문에 데이터가 주어질 때 식(4)의 값을 최대화하는 k 클래스로 출력 변수를 분류할 수 있고, 식(5)와 같이 표현할 수 있다.

$$\hat{y} = argmax_{k\in(1,\ldots,k)} p(C_k)\Pi_{i=1}^{n} p(x_i|C_k) \tag{5}$$

MNB는 모든 변수들이 독립이라는 가정을 사용하기 때문에 데이터 셋에 따라 부정확한 결과를 내는 단점이 있지만, 이 가정을 바탕으로 클래스의 분포를 1차원의 분포로 추정할 수 있게 되어 데이터 셋이 커져도 계산 비용이 적고, 모델의 가정에 따른 결함에도 불구하고 분류 성능이 좋다는 장점이 있다.

3.3.4 Support Vector Machine

SVM은 데이터를 벡터로 표현하여 벡터 공간에 나타내고, 이 벡터 공간을 하이퍼 플레인(hyper-plane)을 기준으로 분류하는 방법을 말하고, 하이퍼 플레인은 식 (6)과 같이 표현된다.

$$w^T x + b = 0 \tag{6}$$

이 하이퍼 플레인은 데이터 벡터들 중 하이퍼 플레인과의 가장 가까운 입력 변수의 사이의 거리를 최대화하는 방법으로 찾아 준다. 그림 7과 같이 중간의 실선은 최적의 경계선(optimal boundary)이고, 이 하이퍼 플레인과 가장 가까운 입력변수를 서포트 벡터(support vectors)라고 한다.

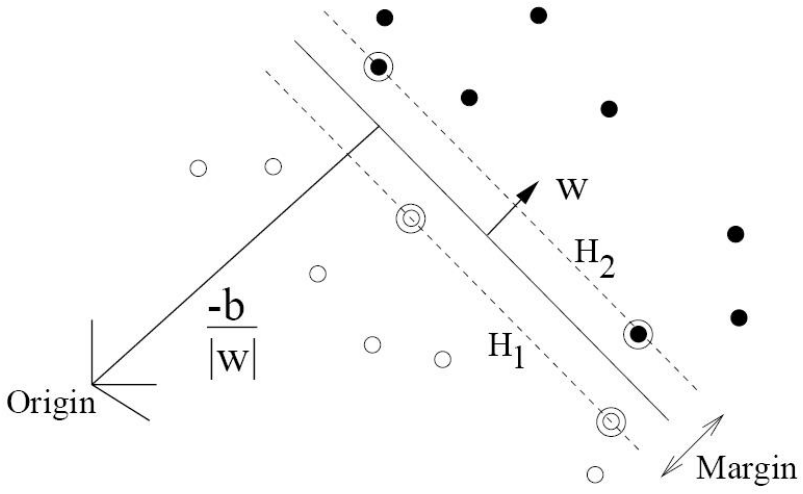


그림 7: Support Vector Machine Framework

이 서포트 벡터와 하이퍼 플레인 사이의 거리를 $1/w$로 나타낼 수 있고, $2/w$를 마진(margin)이라고 부르며, 마진을 최대화하는 하이퍼 플레인을 찾는 최적화 문제를 식 (7)과 같이 표현할 수

있다.

$$\begin{aligned} \min_{w} \quad & w^T w \\ \text{subject to} \quad & y_i(w^T x_i + b) \geq 1 \end{aligned} \tag{7}$$

SVM은 구조적인 위험을 최소화하는 방법을 통해서 오버피팅하지 않으며 우수한 판별 모델을 만들 수 있다는 장점이 있다. 그러나 데이터가 커지면 모델 학습 비용이 커지게 되는 단점이 있다.

3.3.5 SVM with NB features -NBSVM

NBSVM은 SVM과 같은 모델을 사용하는 데, 입력 변수의 형태를 변형하여 사용한다. 먼저 $x_i = r \circ f$ 를 사용하는데, 여기서 r 은 로그 카운트 비율(log count ratio)를 의미하고 그 식은 다음과 같다.

$$r = log\left(\frac{p/||p||_1}{q/||q||_1}\right) \tag{8}$$

위의 식에서 $\mathbf{p} = \alpha + \Sigma_{i:y_i=1} f_i$, $\mathbf{q} = \alpha + \Sigma_{i:y_i=-1} f_i$이고, α는 smoothing parameter이다. 또한 interpolation parameter(β)를 이용한 w'를 다음 식과 같이 계산하여 모델에 사용했다.

$$w' = (1-\beta)\bar{w} + \beta w \tag{9}$$

여기서 $\bar{w}$는 w의 평균 크기(magnitude)이고, $\bar{w} = ||w||_1/|V|$로 계산하여 구할 수 있다. 위의 두 식을 이용하여 SVM을 사용하게 되면 알고리즘의 성능이 향상될 수 있다.

4 Experiments and results

4.1 Data descriptions

텍스트 데이터를 금융 시장에 적용하기 위해 미국 S&P의 공시 데이터인 8-K financial reports[1]를 이용하여 주가 예측을 하고자 한다. 데이터는 id, 시간, 중요한 비지니스 이벤트(bankruptcies, lay-offs, the election of a director, a change in credit, etc), 공시 전문 등으로 이루어져 있다. 데이터의 수집 기간은 2002 ~ 2012년이며, 4개의 기업에 대한 공시 데이터를 수집했고 4개의 기업 리스트는 표 1과 같다.

이 문서에 대한 개별 주식 가격은 Yahoo! Finance[2] 에서 확보하고, 이 개별 주식 가격을 바탕으로 출력 변수 처리를 했다. 주가의 전일 대비 종가(close price)를 두 개의 클래스(UP, DOWN)로 나타내어 출력 변수로 사용했다. 이 때, 개별 종목의 종가의 변동에서 S&P 500 지수의 변동을 빼서 정규화했다. 예를 들어, 개별 종목이 전일 대비 3% 상승을 했고, S&P 500 지수가 전일 대비 1.5% 상승을 했을 경우, 정규화된 변동은 1.5%이게 된다. 이를 바탕으로 정규화된 변동이 1%보다 크면 'UP', 정규화된 변동이 1%보다 작으면 'DOWN'으로 나타내었다.

[1] http://nlp.stanford.edu/∼sidaw/home/

[2] http://finance.yahoo.com/

표 1: Company lists

Ticker Symbol	Company Name	# Doc
C	Citigroup Inc	513
WFC	Wells Fargo & Co	427
GS	Goldman Sachs Group Inc	257
JPM	JP Morgan Chase & Co	835
# Total Doc		**2,032**

각 공시의 문장을 구분하여 각 문장마다 공시 전날의 종가와 공시 후의 종가를 이용하여 두 개의 클래스로 라벨링(labeling)을 했다. 그 결과 UP, DOWN에 해당하는 문장의 개수를 표 2에 표현하여 나타냈고, 이후 절에서는 UP은 positive, DOWN은 negative로 표현하여 사용하도록 한다.

표 2: The number of sentences

	positive	**negative**
# of sentences	25,476,446	27,114,006

4.2 Experiments settings

이 연구에서는 사용하는 텍스트 데이터는 많은 전처리가 필요하다. 그러나 복잡한 전처리를 하지 않고 간단한 stopwords를 제거하고, 다양한 숫자를 'num'으로 바꾸는 등 간단한 전처리만을 통해 연구에 사용했다. 또한 기존 논문 결과와 비교하기 위해 기존 논문에서 사용한 파라미터를 그대로 사용했다. 사용한 파라미터의 값은 표 3에 나타내었다.

표 3: Parameters

Mehtods	**SVM**	**NBSVM**		
Parameters	C	C	α	β
Value	0.1	1	1	0.25

C는 training error와 flatness 사이의 tradeoff를 결정하는 파라미터이며, α는 NBSVM에서 사용되는 smoothing parameter이고, β는 NBSVM의 interpolation parameter를 의미한다. 또한 성능을 평가하기 위해서 10-fold cross-validation을 사용했다.

4.3 Prediction of price percent change after 8-K report announcement

공시가 뜨고 나서 주가의 변화를 살펴보는 실험을 수행했다. 데이터가 커짐에 따라서 SVM의 학습 시간이 오래 걸려서, 비교적 학습 시간이 짧은 MNB를 이용하여 공시와 주가의 변동성을 살펴보

았다. 여기서 MNB를 사용할 때 단어를 unigram, bigram으로 두 가지 방법으로 파싱했다. 공시가 뜨고 나서 1 ∼ 5일 뒤의 주가의 UP, DOWN을 각각 예측한 결과가 표 4와 같다.

표 4: Prediction accuracy according to # days

	1 day	**2 days**	**3 days**	**4 days**	**5 days**
MNB-Uni	**60.95**	59.11	58.62	59.41	58.97
MNB-Bi	**62.63**	58.57	60.39	62.43	61.84

위의 표의 결과와 같이 공시가 뜨고 난 다음날의 주가의 변동이 비교적 예측이 잘 되는 것을 확인 할 수 있었다. 뿐만 아니라 unigram보다 bigram의 성능이 더 좋은 것을 확인할 수 있었다. 이 결과를 바탕으로 공시가 뜨고 난 다음날(1 day)의 주가의 변동에 가장 영향을 많이 준다고 가정하고, 1 day에 대해서 모든 알고리즘을 적용하여 그 성능을 확인했고, 그 결과는 표 5와 같다.

표 5: Prediction accuracy

Methods	**Accuracy**
Unigram(LR)	54.65
Unigram(RF)	59.21
MNB-Uni	62.33
MNB-Bi	64.91
SVM-Uni	63.41
SVM-Bi	63.80
NBSVM-Uni	62.87
NBSVM-Bi	65.67
Distributed Representation	**68.54**

Unigram으로 term feature를 그대로 사용할 경우 LR보다 RF의 성능이 훨씬 좋게 나타나는 것을 확인 할 수 있다. 또한 MNB보다는 SVM의 성능이 더 좋고, NB feature를 SVM에 넣은 방법이 가장 성능이 좋다. 표 4와 마찬가지로 unigram보다 bigram의 성능이 더 좋게 나타난다. 전체적으로 성능이 가장 좋은 방법은 Distributed Representation 방법을 사용하여 단어의 주변을 함께 고려한 경우가 가장 성능이 좋은 것으로 확인 되었다. 뿐만 아니라 Distributed Representation의 경우 computation 시간도 매우 짧아서 데이터 크기가 커지더라도 좋은 성능을 빠른 시간에 얻을 수 있을 것으로 생각된다.

4.4 Sentiment of 8-K report announcement and visualization of stock price

금융 시장은 다른 도메인과 달리 각 문서가 독립적이지 않고, 그 영향이 직간접적으로 주식 가격에 영향을 미친다. 이러한 특성을 바탕으로 공시의 감성을 시각화해보고, 이 감성과 주식 가격의 연관성을 살펴보았다. 4개의 기업 중에 WFC에 대해서 금융위기가 있었던 2008 ~ 2009년의 2년 간의 주식 가격 시각화한 결과가 그림 8와 같다.

그림 8: Stock price of WFC

위의 그래프를 보면 2008년 연초에 주가가 하락하다가 2008년 하반기에 다시 주가가 오르기 시작했고, 2008년 09월을 기점으로 전 세계적인 금융위기와 함께 WFC의 주가가 크게 하락했다. 그 후 2009년 다시 주가가 회복세를 띄었고, 2009년 하반기 이후로 안정세를 찾은 형태로 주가 흐름이 이어졌다. 이 기간 동안 공시의 감성를 시각화 했고, 공시 개수의 한계로 2달씩 묶어서 시각화를 수행했고, 그림 9, 그림 10와 같다.

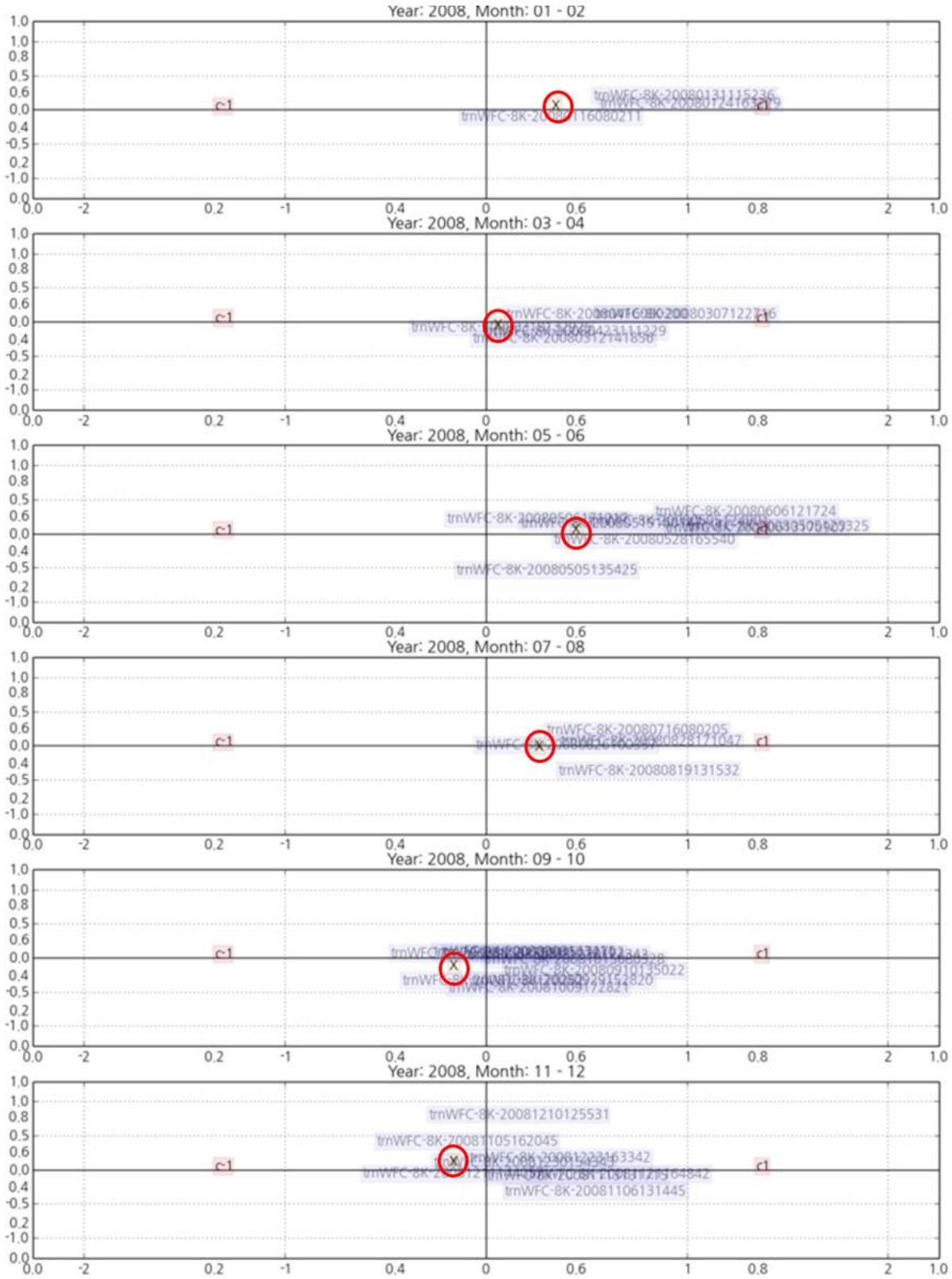


그림 9: Sentiment of 8-K report Announcement (2008)

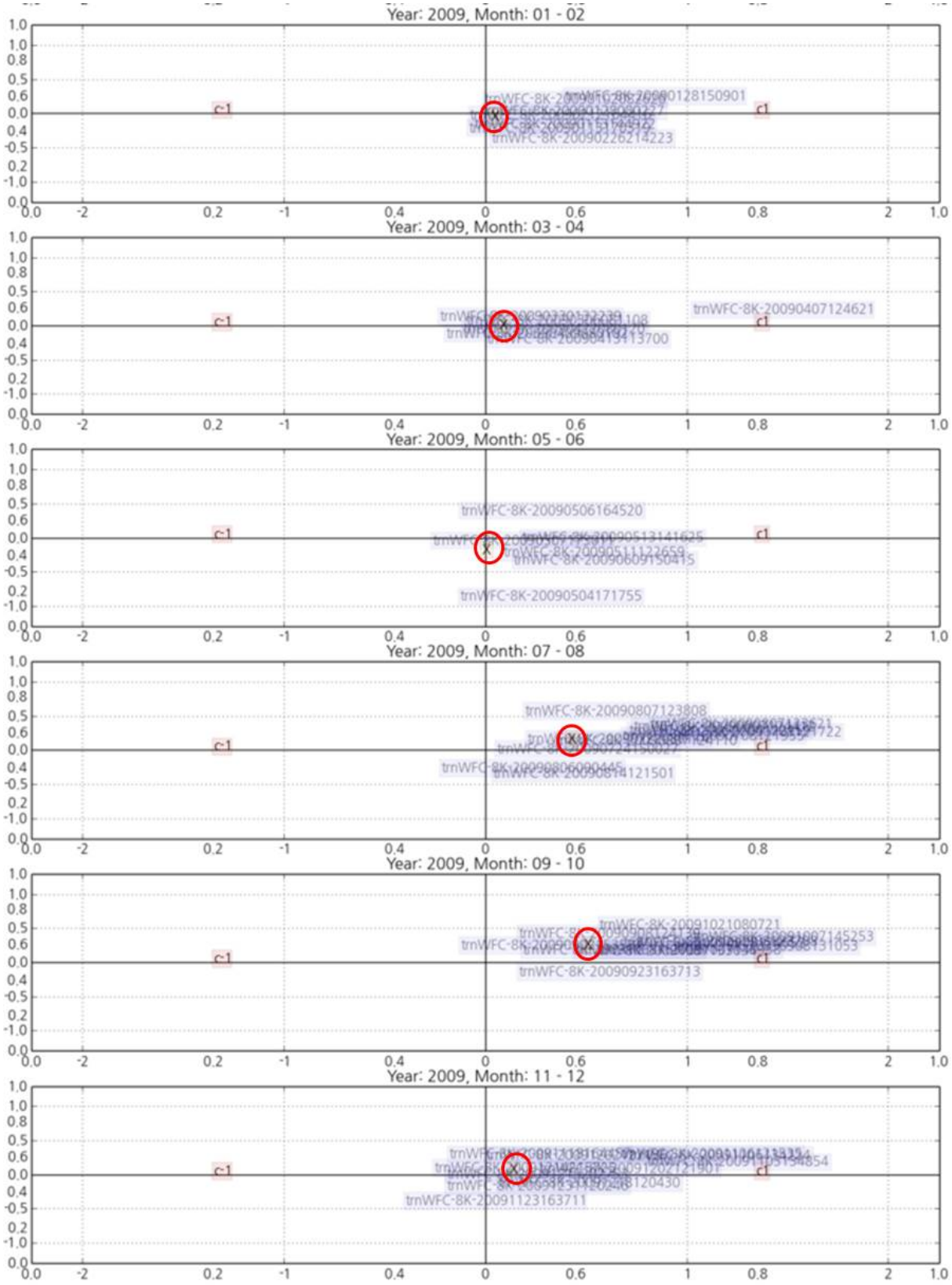


그림 10: Sentiment of 8-K report Announcement (2009)

3.2절에서 설명한 것처럼 위의 그래프에서 x축에 있는 'c-1', 'c1'은 클래스 정보를 공간에 표현한 것으로 'c-1'은 negative 클래스이고, 'c1'은 positive 클래스이다. 파란색으로 표시된 'trnWFC-8K-XXXXX'는 각 기간 동안의 공시 문서의 감성을 표시했다. 마지막으로 'X'는 각 기간 동안 발표된 공시 문서들의 평균 감성을 공간에 표현한 것이다. [Figure 2]의 주가 흐름과 비슷하게 2008년 상반기의 감성이 negative하다가 05 ~ 06월을 기점으로 positive한 공시가 많이 발표된 경향을 나타낸 것을 확인 할 수 있다. 뿐만 아니라 09월 이후로는 급격하게 negative한 공시들이 많이 발표되었다. 2009년도 상반기에도 전 금융위기의 영향으로 인해 공시가 negative한 형태로 나타나다가 하반기가 되면서 positive한 흐름으로 이어지면서 안정적인 형태를 나타낸 것을 확인 할 수 있다.

4.5 Visualization: sentiment of 8-K report announcement of each company

4.4절에서는 하나의 기업에 대해서 기간별로 감성의 변화를 살펴보았고, 이번 절에서는 기업별로 전체 공시에 대한 감성을 시각화 해보았고, 그 결과가 그림 11와 같다.

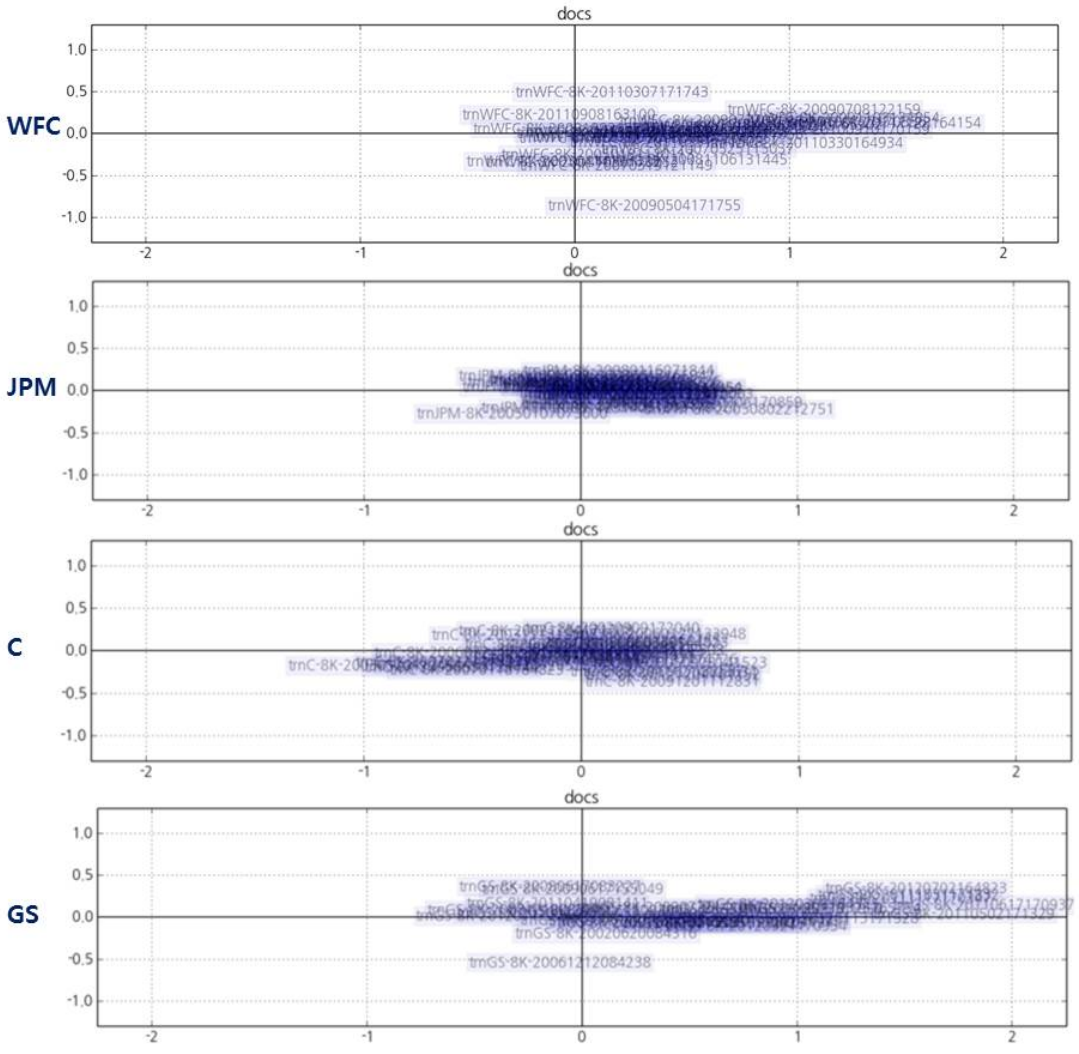


그림 11: Total Sentiment of Each Company

위의 그림에서 보이는 것과 같이 기업 별로 감성의 경향이 다르게 나타나는 것을 확인 할 수 있다. WFC의 경우에는 positive한 공시들이 많은 것을 알 수 있고, JPM의 경우는 전반적으로 neutral한 공시들이 많은 것을 확인 할 수 있다. 반면 C의 경우에는 neutral에 공시들이 많이 있지만 상대적으로 negative한 감성을 갖는 공시가 많이 발표된 것으로 추정할 수 있다. 마지막으로 GS의 경우에는 positive한 감성을 갖는 공시의 비율이 가장 높은 것을 확인 할 수 있다. 이 결과를 바탕으로 기업 C에 대한 주가를 확인해 보았고, 다음과 같은 주가 흐름이 과거에 있었음을 알 수 있었다.

그림 12: Stock Price of Citigroup Inc.

그림 12와 같이 기업 C는 2008 ~ 2009년 금융 위기 때 주가 지속적으로 폭락하여 현재까지 주가 흐름이 회복되지 않고 있는 것으로 확인되었다.

5 Conclusion

본 연구에서는 미국의 4개의 기업에 대한 공시 문서 데이터를 이용하여 주가 방향을 예측해 보았다. 사용한 입력 변수들은 공시 문서의 텍스트 데이터를 이용했고, 도메인이 동일한 4개의 기업의 주가 방향을 UP, DOWN으로 두 개의 클래스 분류 문제로 하여 해결해 보았다. 본 연구에서는 다양한 모델을 기반으로 주가를 예측하는 모델 기반의 주가 방향 예측 방법과 시각화를 기반으로 주가 방향 예측하는 방법으로 두 가지 접근 방법을 시도했다. 모델 기반의 주가 예측을 하기 위해 사용한 분류 알고리즘은 가장 기본적으로 텍스트를 unigram으로 단어를 파싱하여 LR, RF 모델의 입력 변수로 사용하여 예측모델로 활용했고, 텍스트 분류 시 베이스 라인으로 가장 많이 사용되는 MNB, SVM, NBSVM 알고리즘에 대해서 unigram, bigram 두 가지 방법으로 입력 변수를 처리하여 예측 모델에 활용했다. 마지막으로 분산 표상 방법론을 이용하여 주가의 방향을 예측해보았다. 그 결과 분산 표상 방법론은 다른 모델에 비해 정확도도 가장 높고, 학습하는데 시간이 가장 짧게 걸리는 것을

확인 할 수 있었다. 뿐만 아니라 분산 표상 방법 기반으로 감성 분석을 하게 되면 클래스 정보와 단어, 문서를 같은 공간으로 전사시킴에 따라 문서의 감성을 시각화가 가능하다. 이러한 시각화 장점을 이용하여 기간에 따른 특정 기업(WFC)의 감성 변화를 확인할 수 있는데, 금융 위기 시기를 기점으로 2008 2009년의 공시 발표와 주가 흐름의 관계를 살펴 보았다. 그 결과, 주가 흐름이 상승세를 나타낼 때는 공시 발표의 감성이 positive한 것을 알 수 있었고, 주가 흐름이 하락세를 나타낼 때는 공시 발표의 감성이 negative한 것을 알 수 있었고, 그림으로 그 결과를 정성적으로 확인할 수 있었다. 또한 이 때 각 기간별로 평균 등의 대표값을 찾을 수 있고, 이 값을 바탕으로 그 기간의 감성 정도를 지수화 할 수 있었다. 또한 각 기업 별로 전체 문서를 시각화 해봄으로써 감성이 각기 다르게 나타나는 것을 확인 할 수 있었다. 기업 C의 경우 다른 기업들과 다르게 감성이 negative한 문서가 많았는데, 이를 주가 흐름과 비교해보았을 때 주가가 2008년 전후로 크게 하락한 것을 확인할 수 있었다. 이처럼 모델 기반으로 주가 예측을 하게 되면 예측 정확도가 정량적으로 구해지면서 현업의 실무자들에게 정보를 제공해 줄 수 있을 것으로 기대할 수 있고, 또한 시각화 기반으로 주가 예측을 하게 되면 시각화를 통해 주가의 흐름을 한 눈에 보여 줄 수 있어 경영자들에게 의사 결정 지원이 가능할 것으로 기대된다. 이처럼 분산 표상 기법을 활용하여 정량적 및 정성적인 주가 방향 예측을 통해 금융 시장의 다양한 통찰(insight)를 얻을 수 있을 것이다.

또한 본 연구에서는 공시 데이터에서 특정 기업 4개를 샘플링해서 공시 데이터와 주가와의 관계를 살펴보았다. 그러나 일반적으로 학습 데이터 양을 증가시킴에 따라 모델의 예측 성능이 높아질 것으로 예상되기 때문에 특정 도메인을 선택하여 더 많은 데이터를 이용하여 정확도를 구해보고, 그 성능차이를 확인해보는 연구가 필요하다. 뿐만 아니라 단어와 클래스 정보를 같은 공간으로 전사시킴으로써 금융 도메인에 국한된 감성을 나타내는 단어들을 추출할 수 있어, 감성 사전(sentiment dictionary)를 구축하는 데 활용될 수 있는 연구로 발전될 수 있을 것이다.

참고문헌

[1] 인공지능: 2 단계 하이브리드 주가 예측 모델: 공적분 검정과 인공 신경망.

[2] Adebiyi Ariyo, Adewumi O Adewumi, Charles K Ayo, et al. Stock price prediction using the arima model. In *Proceedings of the IEEE International Conference on Computer Modelling and Simulation*, pages 106–112, 2014.

[3] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.

[4] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[5] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167, 2008.

[6] Michael AH Dempster, Tom W Payne, Yazann Romahi, and Giles WP Thompson. Computational learning techniques for intraday fx trading using popular technical indicators. *IEEE Transactions on neural networks*, 12(4):744–754, 2001.

[7] Eugene F Fama et al. *Multiperiod Consumption Investment Decisions*. World Scientific, 1968.

[8] Joumana Ghosn and Yoshua Bengio. Multi-task learning for stock selection. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 946–952, 1997.

[9] Quoc V Le and Tomas Mikolov. Distributed representations of sentences and documents. *arXiv preprint arXiv:1405.4053*, 2014.

[10] Heeyoung Lee, Mihai Surdeanu, Bill MacCartney, and Dan Jurafsky. On the importance of text analysis for stock price prediction. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, 2014.

[11] Shuming Liu. Investor sentiment and stock market liquidity. *Journal of Behavioral Finance*, 16(1):51–67, 2015.

[12] Andrew L Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, volume 1.

[13] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the Advances in neural information processing systems*, pages 3111–3119, 2013.

[14] Tetsuji Nakagawa, Kentaro Inui, and Sadao Kurohashi. Dependency tree-based sentiment classification using crfs with hidden variables. In *Proceedings of the Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 786–794, 2010.

[15] Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing*, volume 1631, page 1642, 2013.

[16] Sida Wang and Christopher D Manning. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers*, volume 2, pages 90–94. Association for Computational Linguistics, 2012.