Journal of the Korean Society of Marine Engineering, Vol. 39, No. 3 pp. 274~279, 2015
J. Korean Soc. of Marine Engineering (JKOSME)
http://dx.doi.org/10.5916/jkosme.2015.39.3.274

ISSN 2234-7925 (Print)
ISSN 2234-8352 (Online)
Original Paper

스탠포드 감성 트리 말뭉치를 이용한 감성 분류 시스템

이성욱†

(Received December 9, 2014 ; Revised January 6, 2015 ; Accepted February 9, 2015)

Sentiment Analysis System Using Stanford Sentiment Treebank

Songwook Lee†

요약: 본 연구는 스탠포드 감성 트리 말뭉치를 이용하여 감성 분류 시스템을 구현하였으며, 분류기로는 지지벡터기계(Support Vector Machines)를 이용하여 긍정, 중립, 부정 등의 3가지 감성으로 분류하였다. 먼저 감성 문장의 품사를 부착한 후 의존구조를 부착하였다. 트리 말뭉치의 모든 노드와 감성 태그를 자동으로 추출하여 문장 레벨의 지지벡터 분류 시스템과 노드 레벨의 지지벡터 분류 시스템을 각각 구현하였다. 자질로는 어휘, 품사, 감성어휘, 의존관계, 형제관계 등 다양한 자질의 조합을 이용하였다. 평가 말뭉치를 이용하여 3클래스로 분류한 결과, 노드 단위에서는 74.2%, 문장 단위에서는 67.0%의 정확도를 얻었으나 2클래스 분류에서는 현재 알려진 최고의 시스템에 어느 정도 필적하는 성능을 거두었다.

주제어: 감성 분류, 지지벡터기계, 스탠포드 감성 트리 말뭉치

Abstract: The main goal of this research is to build a sentiment analysis system which automatically determines user opinions of the Stanford Sentiment Treebank in terms of three sentiments such as positive, negative, and neutral. Firstly, sentiment sentences are POS tagged and parsed to dependency structures. All nodes of the Treebank and their polarities are automatically extracted from the Treebank. We train two Support Vector Machines models. One is for a node level classification and the other is for a sentence level. We have tried various type of features such as word lexicons, POS tags, Sentiment lexicons, head-modifier relations, and sibling relations. Though we acquired 74.2% in accuracy on the test set for 3 class node level classification and 67.0% for 3 class sentence level classification, our experimental results for 2 class classification are comparable to those of the state of art system using the same corpus.

Keywords: Sentiment analysis, Support vector machines, Stanford sentiment treebank

1. 서 론

감성 분류 문제는 사람들이 상품이나 영화 등에 대해 긍정과 부정 중 어떤 의견을 가지고 있는지 파악하는 문제이다. 특정 사안에 대해 사람들이 가지고 있는 의견을 이용하면 마케팅이나 사회적 문제 등의 해결이나 예측에 사용할 수 있으며, 해양산업에도 주요 현안에 대한 일반인의 의견을 신속히 반영하여 다양한 해양산업 발전에 이용할 수 있을 것이다. 감성 분류 문제는 주로 상품평(review)**[1]-[5]** 도메인이나 마이크로블로그(micro-blog)**[6]-[10]** 도메인에 대해서 많은 연구가 이뤄져 왔다.

상품평 도메인은 마이크로블로그 도메인보다 더 길고 복잡한 문장으로 구성되고 있어 감성 정보의 추출에 비교적 용이하나 복잡한 문장구조를 고려해야 정확한 감성 정보를 파악할 수 있다. 이에 반해 마이크로블로그 도메인은 상품평 등의 다른 도메인에 비해 데이터의 길이와 언어의 쓰임이 다른 특징이 있다**[9]**. 대표적인 마이크로블로그 도메인으로는 트위터가 있으며 트위터에서는 다양한 도메인에 대해 사용자의 의견과 감성이 표출된다. 이런 정보들은 주식시장과 영화 박스오피스 수익을 예측하는데 사용되기도 하였다**[6][7]**. 마이크로블로그 도메인은 짧은 길이의 메시지로부터 감성 정보를 추출해야 하기 때문에 상품평 도메인에 비해 이용 가능한 정보가 불충분한 단점이 있다.

스탠포드 감성 트리 말뭉치**[5]**는 영화 상품평 도메인에서 추출된 11,855문장으로 구성되어 있다. 감성 정보가 구문 트리 구조의 각 노드에 부착되어 있고 감성 정보의 구조적인 분석이 가능하도록 완전한 분석 결과를 제공한다.

† Corresponding Author (ORCID: http://orcid.org/0000-0002-6224-4241): Department of Computer Science and Information Engineering, Korea National University of Transportation, 157, Cheoldoparkmulkwanro, Uiwang-si, Gyeonggido, 437-763, Korea, E-mail: leesw@ut.ac.kr, Tel: 070-8855-1686

그 외 감성 정보가 부착된 215,154개의 고유한 감성 구문 사전도 제공하고 있다.

본 연구의 목적은 스탠포드 감성 트리 말뭉치를 이용하여 사용자의 감성을 긍정, 중립, 부정 등으로 분류하는 감성 분류 시스템을 구현하는 것이다. 2장에서는 관련 연구들을 소개하며 3장에서는 시스템 구조에 대해 살펴보고 4장에서 시스템 구현에 필요한 내용을 설명하고 5장에서 실험 결과를 보고하며 6장에서 결론을 맺는다.

2. 관련 연구

먼저 상품평 도메인에 대해서 수행된 연구를 살펴보자. **[1]**은 상품평 도메인의 문서에서 한국어의 본용언의 의미에 영향을 주는 보조용언을 감성 정보의 강도에 따라 분류한 후, 감성 구문에 보조용언이 나타나는 패턴에 따라 그 구문의 극성을 수정할 수 있도록 연관규칙을 정의하였다. **[2]**는 온라인 상품평에 존재하는 장점과 단점 단락을 추출한 후, 상품의 속성을 나타내는 자질을 추출하고, 이 상품 속성 자질을 포함하는 트라이그램을 이용하여 규칙을 생성하여 감성을 분류하였다. **[3]**은 상품 속성 자질을 명시적 자질로 사용하였고 의존관계와 목적어-술어 관계를 이용하여 추출한 규칙을 암시적 자질로 사용하여 감성이 포함된 구문을 추출하였으며 릴렉세이션 레이블링(relaxation labeling) 기법을 이용하여 그 구문의 감성을 결정하였다. **[4]**는 자동차, 여행, 은행 등의 여러 도메인의 상품평에 대해 형용사, 부사 어휘가 포함된 구문을 추출한 후, 이 추출된 구문과 단어 ‘excellent' 사이의 상호정보량(Mutual Information)과 추출된 구문과 단어 ’poor' 사이의 상호정보량을 비교함으로써 그 구문의 감성을 결정하였다.

[5]는 문장의 구문 관계 사이에서 전이되는 감성 정보를 스탠포드 감성 트리 말뭉치와 재귀적 신경 텐서망(Recursive Neural Tensor Network)을 이용하여 처리하였으며, 스탠포드 감성 트리 말뭉치를 사용한 시스템 중 가장 좋은 성능을 보이고 있다.

마이크로블로그 도메인에서 수행된 연구는 대부분 트위터를 이용하였다. **[6]**는 OpinionFinder 도구와 구글 무드 상태 프로파일(Goole-Profile of Mood States) 도구를 이용하여 트위터에 표출된 대중들의 무드 정보를 추출하였고, 추출된 무드 정보를 매일 주식시장 마감시의 다우존스산업평균지수(Dow Jones Industrial Average)의 등락을 예측하는데 이용하였으며, 예측을 위한 기계학습도구로 자동구성퍼지신경망(Self-organizing Fuzzy Neural Network)을 이용하였다. **[7]**은 영화 개봉일 근처의 트위터를 수집한 후 트위터에 존재하는 영화 관련단어와 감성정보를 이용하여 박스오피스 수익을 예측하였다. **[8]**은 트위터 문서를 주관적(subjective) 문서과 객관적(objective) 문서로 구분한 후 각각의 문서의 감성을 분류하였다. 품사 정보, 부정문 유무, n그램 등을 자질로 사용하였고 주관적 문서와 객관적 문서의 감성이 서로 다른 특성을 지니고 있음을 밝혀냈다. 나이브베이시언(Naive Bayes) 분류기를 사용하였다. **[9]**은 트위터에 사용된 이모티콘 정보를 이용하여 트위터 문서의 감성 정보를 자동으로 부착함으로써 지도학습(supervised learning)에 필요한 데이터를 자동으로 생성하였다. 그 후 여러 가지 기계학습 방법을 이용하여 감성을 분류하였다. **[10]**는 소비자 신뢰도와 정치적 의견에 대한 설문조사 결과와 트위터에 나타난 감성어휘의 빈도 사이의 연관성에 대해 연구하였으며 특정 사안에 대해 80% 이상의 연관성이 있는 것을 알아냈다.

3. 시스템 구조

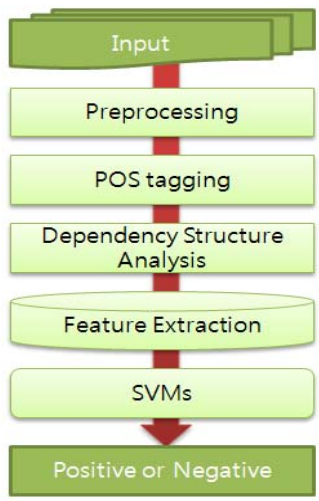

Figure 1: System architecture

Figure 1은 본 연구에서 제안하는 시스템 구조를 나타낸 그림이다. 먼저 스탠포드 감성 트리 말뭉치를 문장 단위 데이터와 노드 단위 데이터로 각각 가공하는 전처리 단계를 거친다. 전처리 단계를 거친 후, 품사 부착 단계와 의존구조 분석 단계를 거쳐 품사 정보와 의존구조 정보를 얻는다. 그 후 자질추출 단계에서는 어휘, 품사, 의존관계, 형제관계 등의 다양한 자질을 추출하여 지지벡터분류기(Support Vector Machines)를 학습한다. 마지막으로 학습된 지지벡터 분류기를 이용하여 평가집합의 입력문장에 대해 긍정, 중립, 부정 등의 감성으로 분류하였다.

4. 감성 분류 시스템 구현

4.1 전처리, 품사태깅 및 의존구조 분석

다음 **Figure 2**는 스탠포드 감성 트리 말뭉치의 한 예와 그 구조에서 추출한 정보를 나타낸 그림이다. 이 말뭉치의 모든 문장은 **Figure 2 (a)**와 같이 이진트리 구조로 분석되어 있으며 각 노드는 0~4까지의 정수값을 가지는데 이 값은 긍정의 정도를 나타낸다(0:매우부정, 1:부정, 2:중립, 3:긍정, 4:매우긍정).

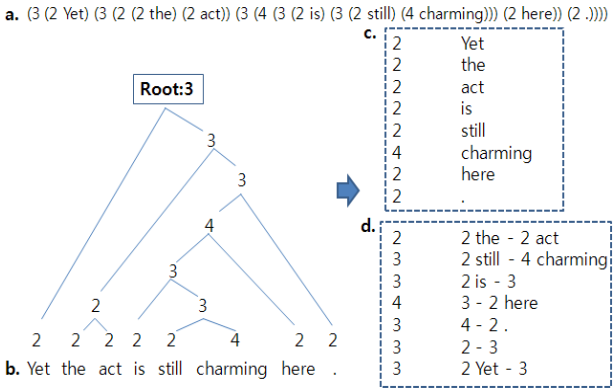

Figure 2: An example of Stanford Sentiment Treebank

전처리기는 **Figure 2 (b)**와 같은 이진트리로부터 단말 노드(**Figure 2 (c)**)와 비단말 노드(**Figure 2 (c)**)를 추출한다. 각 단말 노드와 그 부모 노드의 쌍은 어휘와 그 어휘의 감성 가중치 정보로 이용할 수 있다. 이 쌍을 추후 어휘 자질의 가중치로 이용하기 위해 어휘 사전에 저장한다. 그 외 모든 비단말 노드와 그의 자식 노드 쌍은 노드 단위의 감성 분석을 위한 학습데이터로 사용한다(**Figure 2 (d)**). 이 때, 만약 손자 노드가 단말 노드라면 그 단말 노드의 어휘 정보도 자질로 함께 사용한다. 마지막으로 원시 문장과 루트 노드의 감성태그 쌍을 별도로 추출하여 문장 단위의 감성 분석에 사용한다.

원시 문장의 품사태깅과 의존구조 분석은 스탠포드 Corenlp**[11]** 도구를 이용하였고 출력값은 XML형식으로 출력된다. 다음 **Figure 3**은 예문 “Yet the act is still charming here."에 대한 Corenlp의 품사태깅 결과와 의존구조 분석 결과를 나타낸다.

```
<root>
...
        <tokens>
          <token id="1">
            <word>Yet</word>
            <lemma>yet</lemma>
            <POS>RB</POS>
            <NER>O</NER>
          </token>
            ...
          <token id="7">
            <word>here</word>
            <lemma>here</lemma>
            <POS>RB</POS>
            <NER>O</NER>
          </token>
...
        </tokens>
        <dependencies type="basic-dependencies">
          <dep type="root">
            <governor idx="0">ROOT</governor>
            <dependent idx="6">charming</dependent>
          </dep>
...
<dep type="advmod">
            <governor idx="6">charming</governor>
            <dependent idx="7">here</dependent>
          </dep>
        </dependencies>
```

Figure 3: An example of POS tagging and dependency structure analysis by the Corenlp

4.2 자질 추출

학습을 위한 자질로 어휘, 품사, 의존관계, 형제관계, n그램, 감성어휘 정보, 부정문 등의 다양한 자질의 조합을 사용한다. 의존관계 자질은 수식어-머리어 쌍을 자질로 사용한다. 형제관계 자질은 같은 머리어(head)를 가진 어휘들의 집합을 추출한 후, 모든 가능한 두 개의 단어의 쌍을 조합한 것을 자질로 사용한다. 감성어휘 정보는 **[12]**의 감성어휘사전을 이용한다. 다음은 **Figure 2**의 예문 “Yet the act is still charming here.”을 문장단위 분석을 위해 추출한 자질들의 예이다.

- 어휘: {yet, the, act, is, still, charming, here}
- 품사: {yet/RB, the/DT, act/NN, be/VBZ, still/RB, charming/JJ, here/RB, ./.}
- 의존관계: {(yet->charming), the->act, act->charming, is->charming, still->charming, here->charming}
- 형제관계(yet, act, is, still, here): {(yet, act), ..., (still, here)}
- 바이그램:{(yet the), (the act), ..., (charming here)}
- 감성어휘 정보: {charming}
- 부정문: no

노드 단위 분석에서는 만약 손자노드에 단말노드가 존재한다면 그 단어를 어휘 자질로 사용하고, 또 그 단어와 그 단어의 머리어의 쌍을 의존관계 자질로 사용한다. 손자노드에 만약 비단말노드가 존재한다면, 그 비단말노드의 감성태그만 자질로 사용한다. **Figure 4**는 **Figure 2 (b)**의 루트 노드를 나타낸 것이다.

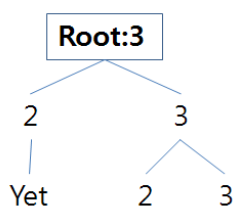

Figure 4: The root node of the example sentence

노드 단위 분석을 위한 **Figure 4**의 루트노드를 위한 자질은 다음과 같다.

- 자식노드의 감성태그: {2,3}
- 어휘:{Yet}
- 품사: {yet/RB}
- 의존관계: {(yet->charming)}

Figure 4의 루트노드의 왼쪽 자식의 자식인 손자노드에 단말노드 “Yet"이 존재하므로 그 단말노드의 어휘정보를

자질로 이용한다. 그러나 오른쪽 손자노드와 같이 단말노드가 존재하지 않으면 오른쪽 자식노드의 감성태그만 자질로 사용한다.

어휘 자질의 가중치로는 3.1절에서 언급한 어휘의 감성태그 정보를 가중치로 이용하고, 그 외 다른 자질들의 가중치는 자질의 존재 유무에 따라 이진 가중치(0 또는 1)로 표현한다.

n개의 자질을 n차원 벡터로 표현하였으며 각 자질의 가중치가 해당 차원의 좌표가 된다. 해당 지지벡터기계 학습은 다중 클래스 분류를 지원하는 Libsvm**[13]**을 사용하였다.

5. 실험

스탠포드 감성 트리 말뭉치는 학습집합(8,544문장), 개발집합(1,101문장), 평가집합(2,210문장)으로 구성되어 있으며, 우리는 학습집합과 개발집합을 합쳐 학습 데이터(9,645문장)로 사용하고, 평가집합을 평가데이터로 사용하였다. 다음 **Table 1**은 자질의 종류에 따른 제안 시스템의 정확도를 나타낸다. 어휘, 품사, 감성어휘 정보, 부정문 등을 기본자질로 사용하였고 기본 자질에 n그램, 의존관계, 형제관계 등의 자질을 추가하여 시스템 성능을 평가하였다.

Table 1: Accuracy of the proposed system for the sentence level with the type of features

Features	3class	2class
Basic features	62.9%	79.3%
Basic+n-gram (2<=n<=3)	67.0%	81.3%
Basic +Dependency Relation	63.8%	80.6%
Basic+n-gram (2<=n<=3) +Dependency Relation	63.9%	80.9%
Basic+n-gram (2<=n<=3) +Dependency Relation +Sibling Relation	52.8%	80.8%

Table 1에서와 같이 기본자질에 바이그램과 트라이그램 자질을 함께 사용하였을 때 3클래스 분류와 2클래스 분류에서 약간의 성능 향상을 보였다. 2클래스 분류는 중립 클래스를 제외한 분류이다. 따라서 2클래스 분류에서는 긍정과 부정으로만 분류한다.

의존관계 자질을 기본자질과 함께 사용하였을 때 3클래스 분류와 2클래스 분류 모두에서 약간의 성능 향상만 얻었고 n그램 자질을 사용하였을 때 보다 성능 향상의 폭이 작았다. 이 때, 평가집합에 사용된 문장의 평균 길이는 19.2개인데 n그램 자질의 경우에는 문장당 평균 12.7개가 사용되었고 의존관계 자질의 경우 문장당 평균 6.8개가 사용되었다. 이는 의존관계 자질의 개수가 n그램 자질의 개수보다 부족함을 의미한다. 또 일부 부정확한 의존관계의 영향도 있었다.

형제관계 자질을 추가하였을 때에는 오히려 약간의 성능 저하를 가져왔으므로 형제관계 자질은 감성분류에 영향을 끼치지 않는다는 것을 알 수 있다.

n그램 자질과 의존관계 자질을 모두 사용하였을 때 성능 향상이 없는 것은 과학습(overfitting)에 기인한다. 다음 **Table 2**를 보자. **Table 2**는 카이제곱통계량을 이용하여 자질을 선택하였을 때 2클래스 분류 시스템의 성능변화를 나타낸다.

Table 2: Accuracy of the proposed system after selecting features with χ^2 statistics

Features	Before	After
Basic+n-gram (2<=n<=3)	81.3%	80.8%
Basic +Dependency Relation	80.6%	82.0%
Basic+n-gram (2<=n<=3) +Dependency Relation	80.9%	81.5%
Basic+n-gram (2<=n<=3) +Dependency Relation +Sibling Relation	80.8%	80.8%

Table 2에서와 같이 2클래스 분류에서 카이제곱통계량을 이용하여 자질을 선택할 경우, 기본자질과 의존관계 자질의 조합에서 약간의 성능향상을 더 얻었다. 이는 좋은 자질을 선택할 때 의존관계 자질이 n그램 자질보다 더 유용하다는 것을 의미한다.

Table 3은 노드 단위에서의 시스템 성능을 나타낸다.

Table 3: Accuracy of the proposed system for the node level with the type of features

Features	5class	3class	2class
Basic features	62.8	73.1	85.4
Basic +Dependency Relation	63.8	74.2	88.3

노드 단위 분류에서는 **Table 3**에서와 같이 기본 자질과 의존관계를 함께 사용하였을 때 더 좋은 성능을 보였다. 노드 단위 분류에서는 노드 안에 있는 정보만 이용하기 위해서 노드를 벗어나는 n그램 자질을 사용하지 않았다.

다음 **Table 4**는 동일한 말뭉치를 사용한 **[5]**의 성능과 제안 시스템을 비교한 표이다. 비교 시스템은 최상의 시스템으로 알려져 있다.

Table 4: Comparison with the state of the art system

	Sentence Level		Node Level	
	5class	2class	5class	2class
RNTN	45.7	85.4	80.7	87.6
The proposed system	39.4	82.0	63.8	88.3

제안 시스템은 비교 시스템에 비해 전반적으로 뒤떨어지는 성능을 보였다. 제안 시스템은 5클래스 분류에서는 비교 시스템보다 성능이 뒤떨어지지만 2클래스 분류에서는 어느 정도 필적할 만한 성능을 거뒀다. 3클래스 분류는 비교 시스템에서는 그 성능이 보고되지 않아 제안 시스템과 비교할 수 없었다. 이러한 성능차이의 원인으로는 감성 구문사전의 사용 유무를 들 수 있다. 비교 시스템은 약 21만개의 감성 구문사전을 말뭉치와 함께 사용하였는데 이 사전에는 말뭉치의 원시문장에 있는 고유 구문과 그 감성정보가 들어 있다. 그러나 감성트리 말뭉치와 달리 학습집합과 평가집합의 구분이 되어있지 않아 제안 시스템에서는 사용할 수 없었다.

6. 결론 및 향후 과제

우리는 스탠포드 감성 트리 말뭉치를 이용한 감성 분류 시스템을 제안하였다. 지지벡터기계를 이용하여 긍정, 중립, 부정 등의 3가지 감성을 학습하였고 학습된 분류기를 이용하여 입력문장의 감성을 분류하였다. 먼저 감성 문장의 품사와 의존구조를 부착하였다. 트리 말뭉치의 모든 노드와 감성 태그를 자동으로 추출하여 문장 단위의 분류 시스템과 노드 단위의 분류 시스템의 구현에 이용하였다. 자질로는 어휘 n그램, 품사, 감성어휘, 의존관계, 형제관계 등 다양한 자질의 조합을 이용하였다. 문장 단위에서는 기본 자질과 어휘 n그램 자질을 사용하였을 때 좋은 결과를 보였으며, 노드 단위에서는 어휘 자질과 의존관계 자질을 사용하였을 때 비교적 좋은 결과를 얻었다. 평가 말뭉치를 이용하여 3클래스로 분류한 결과, 노드 단위에서는 74.2%, 문장 단위에서는 67.0%의 정확도를 얻었으며 2클래스 분류에서는 현재 알려진 최고의 시스템에 어느 정도 필적하는 성능을 거두었다.

향후 노드 단위의 분류기의 성능을 더욱 향상시켜 문장 단위의 분류기의 성능을 개선할 필요가 있으며, 한국어 감성 말뭉치를 구축하여 한국어 감성 분류 연구에 기여할 필요가 있다. 특히 한국어 감성 분류의 경우에도 n그램 자질과 의존관계 자질이 영어와 유사한 결과를 가져올지 실험을 통해 증명할 필요가 있다.

후 기

이 논문은 2013년도 한국교통대학교의 해외파견연구교수지원금을 받아 수행한 연구임.

References

[1] K. J. Lee, “Compositional rules of Korean auxiliary predicates for sentiment analysis,” Journal of the Korean Society of Marine Engineering, vol. 37, no. 3, pp. 291-299, 2013.

[2] B. Liu, M. Hu, and J. Cheng, “Opinion observer : Analyzing and comparing opinions on the web,” Proceedings of the 14th international World Wide Web conference, pp. 342-451, 2005.

[3] A. M. Popescu and O. Etzioni, “Extracting product features and opinions from reviews,” Proceedings of Conference on Empirical Methods on Natural Language Processing, pp. 339-346, 2005.

[4] P. D. Turney, “Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews,” Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02), pp. 417-424, 2002.

[5] R. Socher, A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts Potts, “Recursive deep models for semantic compositionality over a sentiment treebank,” Proceedings of Conference on Empirical Methods on Natural Language Processing, 2013.

[6] J. Bollen, H. Mao, and X. J. Zeng, “Twitter mood predicts the stock market,” Journal of Computational Science, vol. 2, no. 1, pp. 1-8, 2011.

[7] S. Asur and B. A. Huberman, “Predicting the future with social media,” Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, vol. 1, pp. 492-499, 2010.

[8] A. Pak and P. Paroubek, “Twitter as a corpus for sentiment analysis and opinion mining,” Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10), pp. 1320-1326, 2010.

[9] A. Go, R. Bhayani, and L. Huang, “Twitter sentiment classification using distant supervision,” Technical report CS224N, Stanford University, 2009.

[10] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith, “From tweets to polls : Linking text sentiment to public opinion time series,” Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, pp. 122-129, 2010.

[11] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S.

J. Bethard, and D. McClosky, “The stanford coreNLP natural language processing toolkit,” Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics : System Demonstrations, pp. 55-60, 2014.

[12] M. Ganapathibhotla and B. Liu, “Mining opinions in comparative sentences,” Proceedings of the 22nd International Conference on Computational Linguistics, pp. 18-22, 2008.

[13] C. C. Chang and C. J. Lin, “LIBSVM : a library for support vector machines,” ACM Transactions on Intelligent Systems and Technology, vol. 2, no. 3, pp. 27:1-27:27, 2011.