한국컴퓨터정보학회 동계학술대회 논문집 제27권 제1호 (2019. 1)

트위터 감정 분석을 위한 POS 기반의 단어 군집화 기법

김세준[O], 임환희[*], 이병준[*], 김경태[**], 윤희용[*]
[O*]성균관대학교 정보통신대학 전자전기컴퓨터공학과
[**]성균관대학교 소프트웨어대학 소프트웨어학과
e-mail: {ksj105[O], lhh423[*], byungjun[*]}@skku.edu, kyungtaekim76@gmail.com[**], youn7147@skku.edu[*]

Word Clustering Scheme for Twitter Sentiment Analysis Based on POS

Se-Jun Kim[O], Hwan-Hee Lim[*], Byung-Jun Lee[*], Kyung-Tae Kim[**], Hee-Yong Youn[*]
[O*]Dept. of Electrical and Computer Engineering, Sungkyunkwan University
[**]Dept. of Software, Sungkyunkwan University

● 요 약 ●

본 논문에서는 최근 빅데이터 활용 분야의 큰 이슈인 트위터 메시지의 효율적인 감정 분석을 위한 POS 기반의 단어 군집화 기법을 제안하였다. 기존에 군집화를 통한 다양한 텍스트 감정 분석 기법이 제시되어 왔으나, 군집화 된 기능과 분류 결과 간의 관련성에 대한 연구는 미흡하였다. 또한 모든 단어에 대한 감정 분석은 노이즈로 작용될 수 있는 단어로 인해 정확도가 감소할 수 있다. 본 논문에서는 이를 해결하기 위하여 Chi Square 기법을 통하여 분석 결과에 영향을 미치는 단어에 가중치를 부여함으로써 정확도를 향상시킨다.

키워드: 감정분석(Sentiment analysis), 단어 군집화(Word clustering)

I. Introduction

최근 인터넷을 통해 대용량의 데이터가 생성되고 공유됨에 따라 발생하는 데이터에는 다양한 형태가 있으며, 특히 텍스트는 개인 사용자간에 정보를 표현하고 공유하는 데 널리 사용되고 있다. 이로 인하여 텍스트 분류는 학습 데이터를 기반으로 구성된 분석 모델을 사용하여 텍스트 데이터를 카테고리 별로 자동으로 처리하는 이슈가 증가하고 있다. 트위터 API에서 추출한 "트윗(Tweet)"은 감정 분석을 위한 Source 데이터로 적용되어 왔다. 기계학습 기술을 사용한 감정 분석은 트위터 메시지를 '긍정적'또는 '부정적', '중립적'으로 분류한다.

정서 분석을 위해 제안 된 Feature Weighting 기법 중에서 가장 널리 사용되는 것은 특징 빈도와 문서 빈도를 기반으로 한 기법, 단어 빈도 기반 접근 방식인 PSW(Partial Speech-Based Weighting), 등이 있다.

본 논문에서는 식별력이 높은 단어에 가중치를 부여하는 새로운 Feature Weighting 접근법을 제안한다. 제안 된 기법에서 클래스 내 동일한 유형의 Part of speech(POS) 특징을 가지는 단어는 사전에 정의 된 세트로 군집화 된다. 군집화 된 세트와 해당 클래스 간의 종속성은 변형된 Chi Square 기법으로 측정되며 이는 단어들의 차별성과 함께 감정적인 단어들의 가중치를 부여하는 기준으로 사용된다.

II. Preliminaries

1. Related works

1.1 Part of speech

POS tagging은 문장 내 단어들에 명사, 동사, 형용사 등의 Tag를 부여하는 것을 말한다. POS tagging은 텍스트 분류, 음성 인식, 자동 번역 등 다양한 문제를 해결하는데 해결책으로 제시되어 왔으며 POS tagger 또한 Brill tagger, Tree tagger, CLAWS tagger 등 다양하게 존재한다.

POS tagging의 학습 단계에서 말뭉치는 서로 다른 문맥 환경에서 단어를 제공하기 위하여 사용되며 문맥상 정보는 단어의 어휘 클래스를 결정하는데 필요한 규칙을 구성하기 위한 단서로 이용된다. 이를 통하여 문맥에서 단어가 등장할 확률을 계산함으로써 가장 적합한 태그가 선택된다.

III. The Proposed Scheme

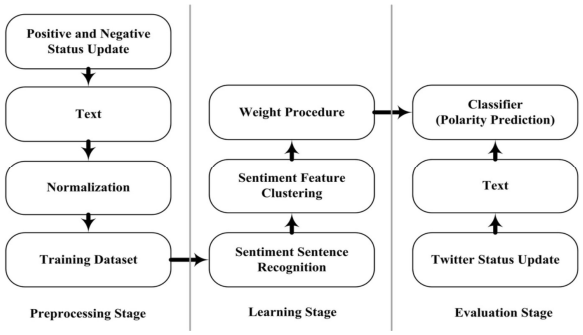

Fig. 1. 제안하는 이주 기법의 구조

제안하는 기법의 전체적인 동작의 흐름은 다음과 같다. 먼저 POS Tagging을 통하여 일부 문장을 기준에 따라 설정된 훈련 데이터로 선택하고, 극성에 따라 긍정 및 부정이라는 두 클래스로 분류한다. 문장의 단어들은 POS 태그를 사용하여 군집화 된다. 이 때 Chi Square를 이용하여 군집화 된 단어가 속해있는 특성 집합의 종속성과 단어의 식별성에 따라 모든 단어에 가중치가 할당된다. 학습 단계가 끝나면 통계 데이터 표가 생성되며 테스트 문서에 있는 트위터 문장들의 감정은 통계표에 근거하여 판단된다.

위 과정에서 POS Tagging은 다음과 같이 진행된다. 먼저 문장 감정 분석은 주로 초기 말뭉치와 문서 $\Phi = \{d_1,...,d_{|\Phi|}\}, d_i = \{s_1,...,s_{|S|}\}$ 와 미리 정의 된 클래스, $C = \{c_1,...,c_{|C|}\}$의 유용성에 따라 결정된다. 여기서 d_i는 초기의 말뭉치와 $|C|$ 클래스의 $|\Phi|$ 문서 중 $|S|$ 문장으로 구성된 문서를 나타낸다. 목적 함수는 분류 과정을 특징짓는 함수 $\Psi : \Phi \times C \rightarrow$ {N, P}로 나타낼 수 있으며, $\Psi(d_i,c_j) = N$이면 $document_d_i$는 c_j의 부정적인 데이터 세트, 반면 $\Psi(d_i,c_j) = P$라면 d_i는 긍정적인 데이터 세트로 정해진다. 초기의 말뭉치 Φ는 D_{neg} 와 D_{pos} $(D_{neg} \cup D_{pos} = \Phi)$ 두 개의 클래스 집합으로 분류되며 $D_{neg} = \{d_1,\cdots,d_{sn}\}$, $Dpos = \{d_1,\cdots,d_{sp}\}, sn+sp = |\Phi|$ 로 정의된다. sn과 sp는 각각 부정과 긍정의 문장으로 구성된 문서의 크기이고, D_{neg}와 D_{pos}의 모든 문장은 POS tagger를 통해 분석되며 문장의 모든 단어에는 해당 POS 태그가 지정된다.

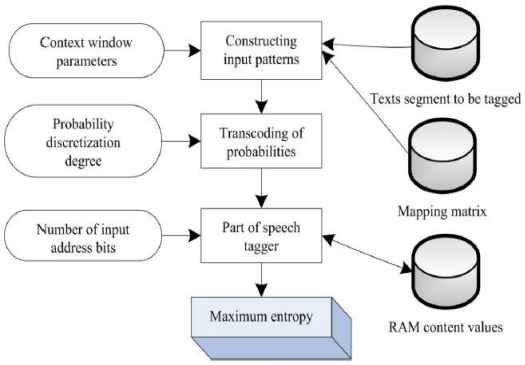

Fig. 2. POS tagging 과정

IV. Conclusions

본 논문에서는 특징 가중치 부여를 이용한 트위터 문장의 감정 분석 기법의 정확도를 향상시키기 위한 방법으로 POS tagging을 이용하였다. 제안하는 기법은 문장을 기능적으로 분할하고 가중치를 적절히 부여함으로써 학습을 통하여 정확도를 높여감으로써 빅데이터 활용의 효과적인 분석을 제공할 것으로 기대된다.

ACKNOWLEDGEMENT

본 연구는 과학기술정보통신부 및 정보통신기술진흥센터의 정보통신·방송연구 개발 사업(No. 2016 -0-00133, 초연결 IoT 노드의 군집 지능화를 통한 Edge Computing 핵심 기술 연구), SW중심대학지원사업(2015-0-00914), 한국연구재단 기초연구사업 (No.2016R1A6A3A11931385, 실시간 공공안전 서비스를 위한 소프트웨어 정의 무선 센서 네트워크 핵심기술 연구, 2017R1A2B2009095, 실시간 스트림 데이터 처리 및 Multi-connectivity를 지원하는 SDN 기반 WSN 핵심 기술 연구), BK21PLUS 사업의 일환으로 수행되었음.

REFERENCES

[1] A. Yang et al., “Enhanced twitter sentiment analysis by using feature selection and combination”, SocialSec 2015, pp.52-57, 2016.