ISBN 978-981-14-1455-8
Proceedings of 2019 the 9th International Workshop on Computer Science and Engineering
WCSE_2019_SPRING
Yangon, Myanmar, February 27-March 1, 2019, pp. 81-85
doi:10.18178/wcse.2019.03.014

Newly-Coined Words and Emoticons Dictionary Construction for Social Data Sentiment Analysis

Jin Sol Yang [1] and Kwang Sik Chung [2 +]

[1] Dept. of Computer Science, Graduate School of Korea National Open University, Dongsung-dong, jongno-gu, Seoul 110-791, Korea

[2] Dept. of Computer Science, Korea National Open University, Dongsung-dong, jongno-gu, Seoul 110-791, Korea

Abstract. In the composition of social networks and social networks services, most of articles are written with writer's personal ideas and emotions within 200 words. Social networks services produce social big data and provide vast volumes of data. Social big data has various types of text-centric data, photos, music and videos, and informal data. It is difficult to deal with social big data, since it has rapid changes and volume growth. But, analysis of social big data may accept the online market trust agreement and establish national policies to support the marketing and strategy formulation of the company for marketing and strategy establishment.

This paper proposes a new social big data analysis system that uses new-coined words and emoticons for social big data analysis. In this paper, we also construct a new dictionary of emoticons to improve the accuracy of big data analysis related to the public opinion, and to improve the accuracy of emotional analysis using the existing emotional dictionaries and the newly constructed new emoticons and emoticon dictionaries at the analysis stage. Especially, the accuracy of social big data analysis could be improved by quantifying and analyzing emotional level about emoticons and new-coined words. This research built a dictionary of emoticons, and used the Naver open dictionary for newly-coined words. But, newly-coined words that were not included in the Naver open dictionary are added to proposed newly-coined words database.

Keywords: Social media opinion, social big data, social big data analysis, emoticon, newly-coined words

1. Introduction

As social networks services become more active, people try to imply their opinions or thoughts on various SNS into less than 200 words. Social big data is a vast amount of informal data produced on social media, the amount of which is increasing exponentially and spreading rapidly. Also, social big data is mainly composed of mixed texts, music, and images are mainly composed of existing text-oriented data. The analysis of the emotions and public opinion of these social big data can be used to create marketing value and strategy of the company, which can generate the emerging value of online market. Also the government recognizes social big data as the data of the public opinion and sentiment, and is actively utilizing it to establish national policies[1,2]. For this reason, researches on text mining, natural language processing, and emotional analysis are being actively carried out. The basic premise of these research fields is to acquire social big data and construct analysis dictionaries to be used for SNS big data analysis. In particular, emotional analysis based on social big data can help people's consumption aspect and product evaluation as well as corporate sales and policy establishment. However, social big data is informal data and contains many new coined words and emoticons. Therefore, there is a limitation on the accuracy and analysis range of emotional analysis. The new coined word contains the social phenomena and trends of modern society implicitly, and the emoticons are electronically produced by letters and symbols, and they express the

+ Corresponding author. Tel.: + 82-2-3668-3654; fax: +82-2-3668-4650.
E-mail address: kchung0825@knou.ac.kr.

emotional state more implicitly and concisely than the general text. Although the coined words and emoticon are an important part of the emotional analysis, they are excluded from the emotional dictionary and analysis.

The remainder of paper is organized as follows. In chapter 2, we review previous text analysis research for social media analysis. The proposed social media analysis system are described in chapter 3. We introduce the social big data architecture and function modules of social big data analysis are defined and described in chapter 3. Finally, we conclude in chapter 4.

2. Related Works

Song Eunji[3] stated that sentences written on online can not be analyzed correctly by the existing morpheme analyzer because there are many grammatical errors and mistakes in spelling and spacing, and the length of sentences is too short to understand the exact meaning. In order to solve these problems, this research used the word selection method using the priority of the words in the sentence. In this paper, we propose an emotional analysis module that constructs a word property database that is dependent on the part-of-speech by separating verb and cognition based on the part-of-speech information extracted from the morpheme analyzer. Song Eunji[3] used SVM algorithm for text emotion analysis. The SVM algorithm is an emotional analysis technique through machine learning, which differs from this paper based on social big data emotion dictionary. In addition, Song Eunji[3] can compensate for errors in spelling and spacing, but emotional analysis is difficult for actual coined words and emoticons. And previous researches for emotional analysis have focus on the opinions of web sites and social media services and provided analysis results of 'positive / negative' or 'good / no' [4,5,6,7].

In this paper, to solve these problems, emotional words information are extracted from the coined word and emoticon by additionally using the coined word and emoticon dictionary that are proposed in this paper.

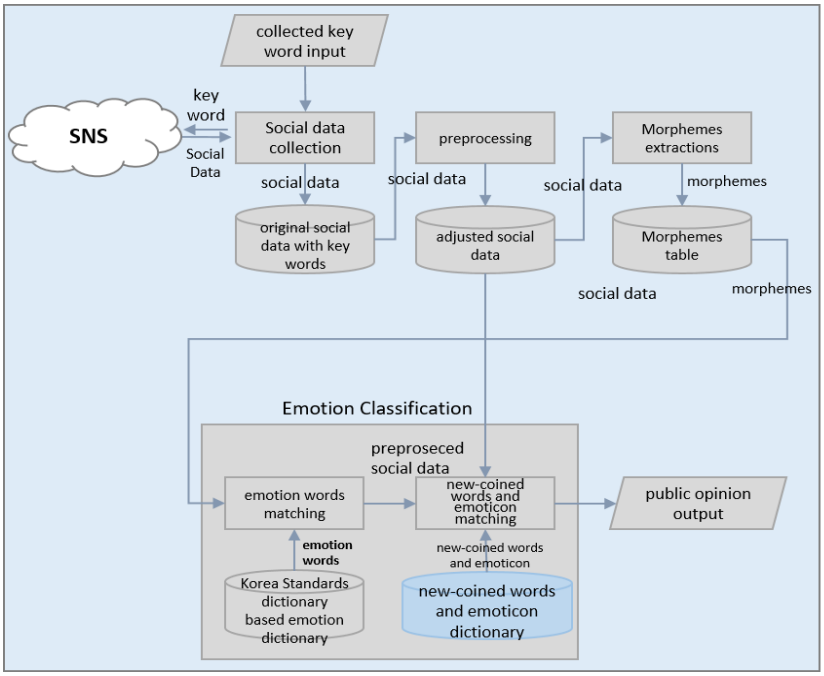

Fig. 1: Extraction Process of Newly-coined words or emoticon

3. Social Big Data Analysis System for Newly-Coined Words and Emoticons

Generally, the Internet news provided by the media uses the standard language as a basis. However, the social data created by an individual includes many new words and emoticons which are non-standard words. The coined words and emoticons are utilized in SNS and have direct emotional meaning. For example, in the social data "The man is really rainbow manners", "rainbow manners" is a new word that combines "ignorance" and "persona" and has a negative meaning. However, when performing emotional analysis based on the existing Korean dictionary, "rainbow manners" is excluded from the analysis even though it is an important emotional word. It is possible to improve the accuracy of emotional analysis by registering the coined words and emoticons in the emotional dictionary based on the existing Korean dictionary and using

them as emotional words. This study collects social data from SNS for the construction of new words and emoticon emotion dictionary with process of figure 1. And extracts the coined words and emoticons from the collected social data. The social data collection method uses the API provided by the SNS company, and the extraction of the coined word and emoticon from the collected social data uses the algorithm proposed in this study.

3.1 Newly-coined words and Emoticon Collect Process

When collecting social data for extracting newly-coined words and emoticons, the Newly-coined words and Emoticon Collect Process randomly collects social data using APIs provided by SNS vendors. The randomly collected social data is stored in a raw data table for extracting newly-coined words and emoticons. The raw data table for extracting the newly-coined words and the emoticon is composed of the original text number (ID) and the content (CONTENT) in figure 2 and figure 3. The original number is an automatically generated key value when collecting social data, and the content means collected social data.

```
public static String ClearTweetID(String content) {
Pattern URL = Pattern.compile("([@+])((\\w+[-a-zA-Z0-9:@;?&=\\/%\\+\\.\\*!'\\(\\),\\$_\\{\\}\\^~\\[\\]`#|]\\s+)");
  Matcher m;
  m = URL.matcher(content);
  content = m.replaceAll("");
  return content;
  }
```

Fig. 2: URL deletion code

```
public static String ClearUrl(String content) {
   Pattern URL = Pattern.compile("((\\w+:\\/\\/)[-a-zA-Z0-9:@;?&=\\/%\\+\\.\\*!'\\(\\),\\$_\\{\\}\\^~\\[\\]`#|]+)");
  Matcher m;
  m = URL.matcher(content);
  content = m.replaceAll("");
  return content;
  }
```

Fig.3: User ID deletion code

(•ε• ;)	(o゜д゜o)	(; ◔д◔)	(´o｀)	(´^益^｀)♥
(◕‿◕✿)	(╋⊙д⊙)	(´Д｀*)	(´▽｀)	๑♥‿♥๑
(。[oo]ˬ[oo]。)♡	(; ◔д◔)	♪(*´∀｀)	(´*｀)	๑•‿•๑
(・´з`・)	(´^益^｀)	♪(´ε｀*)	(´～｀)	o(๑◕‿‿◕๑)o~♪
(´゜з゜｀)	(･益･。)	♪(*´θ｀)ノ	(; ◔д◔)♥	(*´ワ`*)"

Fig. 4: Character emoticon

3.2 Syntactic Word Separation Process

In the syntactic word separation process, the syntactic word tokenizing operation is performed after separating the social data into the syntactic word units. The original social data table for extracting newly-coined words and emoticons stores only some data of the entire social data. A raw social data table for extracting the newly-coined word and the emoticon is extracted by extracting only some data of the entire social data. The boundary between the word and the word is judged based on the spacing. And tokenize the words in spacing units in the original social data table for extracting coined words and emoticons. Then, when performing emotional analysis, the original social data table for emotional analysis including the collected keyword is also tokenized in a space unit.

3.3 Emoticon Analysis Process

In this study, three types of emoticons are extracted from tokenized words. First, it extracts emoticons of the form consisting of a combination of characters such as "ㅅㅅ ㅋㅋ" and "ㅠㅠ ㅠㅠ" in figure 4. Korean character(Hangul) expresses characters by combining the three elements of initial, neutral, and longitudinal. Therefore, Hangul can not express a character as a single element of primitive, neutral, and longitudinal. However, most of the emoticons are composed only of the beginning. Using these characteristics, the word consisting of only the beginning in the tokenized word is judged as emoticons and extracted. The extracted character type emoticons are stored in the emoticon candidate table.

Fig. 5: Image emoticons

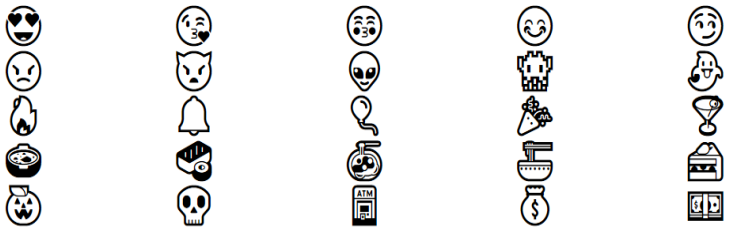

Fig. 6: 4-bytes Character emoticons

Second, we extract emoticon of image form as shown in figure 5. Recently, as the online community evolves, it is changing from a character-shaped emoticon to an image-type emoticon. SNS companies develop their own unique image emoticons and provide them to users. Image-type emoticons are usually composed of <IMG> tags in HTML format. After extracting the <IMG> tag from the tokenized word, it finds unique patterns of SNS companies. For example, if there is a value of "sticker image" in alt attribute in the emoticon tag of image form in figure 5. <IMG> tag would be extracted. The extracted <IMG> tag is stored in the emoticon candidate table like a character type emoticon. Third, extract emoticons in 4-Byte Unicode character form in figure 6. 4-Byte Unicode characters are extended from 4-Byte to 2-Byte Unicode characters, so that it is possible to represent characters in picture format. Emoticons that use 4-Byte Unicode characters are "Emoji" developed by NTT Tokomo Co., Ltd. in Japan. "Emoji" is supported by Apple and Google and by Facebook. A 4-Byte Unicode character type emoticon extraction process such as "Emoji" looks for all 4-Byte Unicode-encoded characters by examining all the characters in the tokenized word. The extracted 4-Byte Unicode character type emoticons are stored in the emoticon candidate table like letter type emoticons.

3.4 Newly-coined Words Analysis Process

The Naver Open Dictionary is used as the basis for constructing a new emotion dictionary. The Naver Open Dictionary, a user-participatory open dictionary, has 32 new languages, including Korean, English, Chinese and Japanese, registered on the website (opendict.naver.com). The user can directly register a new coined word that is not registered, and can ask another user a question about the coined word. The registered coined words are displayed on the page in the "Real-time words". In this study, the search condition of the "real-time word" list is set to Korean. The web crawler then collects words and meanings from the "real-time words" list. The collected words and meanings are stored in the candidate word table. The Naver Open Dictionary is used as the basis for constructing a new emotion dictionary.

3.5 Pre-registration of Newly-coined words and Emoticon

If extraction of the coined word and emoticon is completed, the registrant registers the coined word and the emoticon emotion dictionary. The registrant categorizes the new coined words and emoticons in the coined word candidate table and the emoticon candidate table. Valid new words that are not included in the general emotional dictionary and the coined word dictionary include buzzwords and food items that represent political and social issues. The feeder is a slang word used mainly by young people. Among adults, the frequency of use of food stuffs is increasing, and the use of food stuffs is spreading rapidly also in SNS. Valid emoticons that are not registered in the emoticons dictionary include character emoticons, emoticon emoticons, and image tag emoticons. The classification of valid coined words and emoticons is done

manually. To solve this problem, we developed a. The dedicated tool can search the extracted newly-coined words and emoticons in detail as shown in figure 7, and can specify polarity and weight. Polarity can be classified into positive and negative, and weights can range from 1 to 5. The new coined words and emoticons registered using the dedicated tool are stored in the emotion dictionary table.

4. Conclusion

In SNS, everyday too much mentions and twits would be made and delivered to or by everyone. It is too fast to make new mentions and twits, also. But, volumes and speed of new mentions and twits are not important to understand SNS polarity about a theme, since the high computing power with high-speed CPU and huge main memory can deal with the speed and volume of new mentions and twits. Nowadays, newly-coined words and new emoticons are made with various types and formats so that newly-coined words and new emoticons can easily and simply express people's emotion state. It is hard to analyze and distinguish this kinds of newly-coined words and new emoticons makes.

In this research, newly-coined words and new emoticons dictionary construction methods are proposed. With this kind of newly-coined words and new emoticons dictionary, SNS big data can be analyzed more detailed so that we can make the more specific and accurate SNS big data analysis results. At the first step, the morphological analyzer extracts the morphemes as same as the existing emotional analysis method and extract emotional words by comparing them with the basic emotional dictionary in the extracted morpheme. In the second step, the newly-coined word and emoticon are extracted from the pre-processed data, before the morphological analysis using the newly-coined word and emoticon dictionary.

5. References

[1] kOSAC : Munhyong Kim O Ha-Yeon Jang O Yu-Mi Jo O Hyopil Shin, "KOSAC: Korean Sentiment Analysis Corpus", proceedings of KOREA INFORMATION SCIENCE SOCIETY, 2013.6, pp. 650-652

[2] Phil-Sik Jang , Study on Principal Sentiment Analysis of Social Data, Journal of the Korea Society of Computer and Information 19(12), 2014.12, pp. 49-56

[3] Eun-Jee Song, "The Sensitivity Analysis for Customer Feedback on Social Media", Journal of the Korea Institute of Information and Communication Engineering 19(4), 2015.4, pp. 780-786

[4] Bollen, J., Mao, H., & Zeng, X. Twitter mood predicts the stock market. Journal of computational science, 2(1), 2011, pp.1-8.

[5] DiGrazia, J., McKelvey, K., Bollen, J., & Rojas, F. More tweets, more votes: Social media as a quantitative indicator of political behavior. PloS one, 8(11), 2013, e79449.

[6] Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. Sentiment analysis of twitter data. In Proceedings of the workshop on languages in social media (pp. 30-38). Association for Computational Linguistics., 2011.

[7] Yadollahi, A., Shahraki, A. G., & Zaiane, O. R.. Current State of Text Sentiment Analysis from Opinion to Emotion Mining. ACM Computing Surveys (CSUR), 50(2), 25. 2017