future internet

MDPI

Article

Social Emotional Opinion Decision with Newly Coined Words and Emoticon Polarity of Social Networks Services †

Jin Sol Yang [1], Myung-Sook Ko [2] and Kwang Sik Chung [1,*]

[1] Department of Computer Science, Graduate School, Korea National Open University, Jongno-gu, Dongsung-Dong, Seoul 30387, Korea
[2] Department of Business Administration, Bucheon University, WonMi-Gu, SimGok-Dong 25, Bucheon 14632, Korea
* Correspondence: kchung0825@knou.ac.kr; Tel.: +82-23668-4654

† This paper is an extended version of our paper: Yang, J.S.; Chung, K.S. Newly-Coined Words and Emoticon Polarity for Social Emotional Opinion Decision. In Proceedings of the IEEE 2nd International Conference on Information and Computer Technologies, Kahului, HI, USA, 14–17 March 2019; p. 76.

Received: 9 June 2019; Accepted: 22 July 2019; Published: 25 July 2019

Abstract: Nowadays, based on mobile devices and internet, social network services (SNS) are common trends to everyone. Social opinions as public opinions are very important to the government, company, and a person. Analysis and decision of social polarity of SNS about social happenings, political issues and government policies, or commercial products is very critical to the government, company, and a person. Newly coined words and emoticons on SNS are created every day. Specifically, emoticons are made and sold by a person or companies. Newly coined words are mostly made and used by various kinds of communities. The SNS big data mainly consist of normal text with newly coined words and emoticons so that newly coined words and emoticons analysis is very important to understand the social and public opinions. Social big data is informally made and unstructured, and on social network services, many kinds of newly coined words and various emoticons are made anonymously and unintentionally by people and companies. In the analysis of social data, newly coined words and emoticons limit the guarantee the accuracy of analysis. The newly coined words implicitly contain the social opinions and trends of people. The emotional states of people significantly are expressed by emoticons. Although the newly coined words and emoticons are an important part of the social opinion analysis, they are excluded from the emotional dictionary and social big data analysis. In this research, newly coined words and emoticons are extracted from the raw Twitter's twit messages and analyzed and included in a pre-built dictionary with the polarity and weight of the newly coined words and emoticons. The polarity and weight are calculated for emotional classification. The proposed emotional classification algorithm calculates the weight of polarity (positive or negative) and results in total polarity weight of social opinion. If the total polarity weight of social opinion is more than the pre-fixed threshold value, the twit message is decided as positive. If it is less than the pre-fixed threshold value, the twit message is decided as negative and the other values mean neutral opinion. The accuracy of the social big data analysis result is improved by quantifying and analyzing emoticons and newly coined words.

Keywords: social media opinion; social big data; social big data analysis; emoticon; newly coined words; social emotional opinion

1. Introduction

As social networks services become popular and expresses social opinions more efficiently, social network service participants increase, and this increases the volume of social big data. Social big data increases a vast amount and this is increasing exponentially and spreading rapidly on social media. Also, social big data are mainly composed of mixed texts, music, and are the existing text-oriented data. By analyzing the emotions, opinions, and public opinion of these social big data, companies can create an emerging online market when social big data are used for corporate marketing and strategy development. The Korean government and global Korean company recognize social big data as important data and actively utilize it for customer's needs and demands. For this reason, research on text mining, natural language processing, and emotional analysis is actively carried out. The emotional analysis based on social big data can help people's consumption aspect and opinions on government policy or company product evaluation. According to the research [1], companies analyze and survey consumers' opinion not only for their images, but also for other companies' products and services. However, social big data contains many newly coined words and emoticons. The newly coined words and emoticon express people's emotions and opinions appropriately and are directly related to the social emotion. These newly coined words and emoticons are difficult to classify systematically based on morphemes or are difficult to be preregistered. In this paper, Korean-based newly coined words and global emoticons are extracted from Twitter. The emotional polarity of SNS with the coined words are analyzed based on it.

The remainder of the paper is organized as follows. In Section 2, we review previous emotional analysis research and show weak points of them. The proposed social emotional opinion decision algorithms with newly coined words and the emoticon polarity of social network services (SNS) are described in Section 3. We introduce the architecture and function processes of social emotional opinion decision with newly coined words and emoticon dictionary and the experiment results are described in Section 4. Finally, we conclude in Section 5.

2. Related Works

Most of the early emotional analysis studies have analyzed the opinions of websites and social media services and provided analysis results of 'positive/negative' or 'good/no' [2–8]. In [9], online blog sentences that cannot be correctly analyzed by the previous morpheme analyzer is focused on, because they have many kinds of grammatical mistakes and misspellings. Further, 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, an emotional analysis module that constructs a word property database is proposed. The emotional analysis module is dependent on the part of speech by separating verb and cognition based on the part-of-speech information extracted from the morpheme analyzer. Ref. [9] used Support-vector machine (SVM) algorithm for text emotion analysis. Ref. [9] can compensate for errors in spelling and spacing, but emotional analysis is difficult for actual newly coined words and emoticons in [9]. In order to solve these problems, emotional words and information are extracted from the newly coined word and emoticon by additionally using the coined word and emoticon dictionary that are proposed in this paper. In [10,11], polarity is defined only as negative and positive, and did not considered newly coined words and emoticons. In [12–14], polarity is classified between negative and positive. However, formal words and a formal dictionary are used to analyze SNS opinions. Previous Korean emotion classification research has not yet established a formal classification system for Korean emotion categories, and Korean emotional resource such as SentiWordNet in English is also lacking. This leads to the absence of learning corpus used in the machine learning approach, suggesting the difficulty of classifying Korean emotions.

Previous Korean emotional analysis techniques exclude profanity, abbreviation, newly coined words and emoticon, and only analyzing standard vocabulary in dictionaries. However, many articles (texts) such as SNS, homepage comments, and blogs contain spoken words, idiomatic expressions,

newly coined words, and emoticons. By filtering these texts, spoken words, idiomatic expressions, newly coined words, and emoticons are excluded from the emotional analysis process. In particular, Korean characters have vowels and consonants so that newly coined words come from a combination of vowels and consonants. These kinds of newly coined words need different emotional analyses from English and have words that are constructed from English. In this paper, emotional analysis methods for Korean-based newly coined words and emoticons are proposed.

3. Newly Coined Words and Emoticon Extraction

Generally speaking, the internet news provided by mass media uses standard words. However, the social data created by individuals include many non-standard words (newly coined words, emoticons, etc.). Newly coined words and emoticons are utilized in SNS and they have sentimental meaning. For example, the 'mujigaemaeneo (very poor manners)' in the article "geu salam-eun jeongmal mujigaemaeneoya!!! (He is in very poor manners!!!)' in social data is a newly coined word that combines 'muji (very)' and 'gaemaeneo (poor manners)' and has a negative meaning. However, when performing the emotional analysis based on the existing Korean dictionary, "mujigaemae (very poor manners)" of "geu salam-eun jeongmal mujigaemaeneoya!!! (He is in very poor manners!!!)" are excluded from the analysis even though they are important emotion words. If we separately set up newly coined words and emoticons as sentimental dictionaries and use existing emotional dictionaries based on Korean dictionaries, we can improve accuracy in emotional analysis. This study uses Twitter and Naver Blog to establish newly coined words and emoticons-based sentimental dictionaries. From collected social data, Twitter extracts the newly coined words, character type emoticons, 4-byte character type emoticons, and Naver blog extracts image type emoticons. Twitter uses its own search application programming interface (API) to collect social data, and since Naver does not provide its own API, it develops a separate crawler and collects blogs. We then extract the newly coined words and emoticons from the collected social data using the proposed algorithm. Figure 1 shows the whole process of the newly coined words and emoticon extraction process.

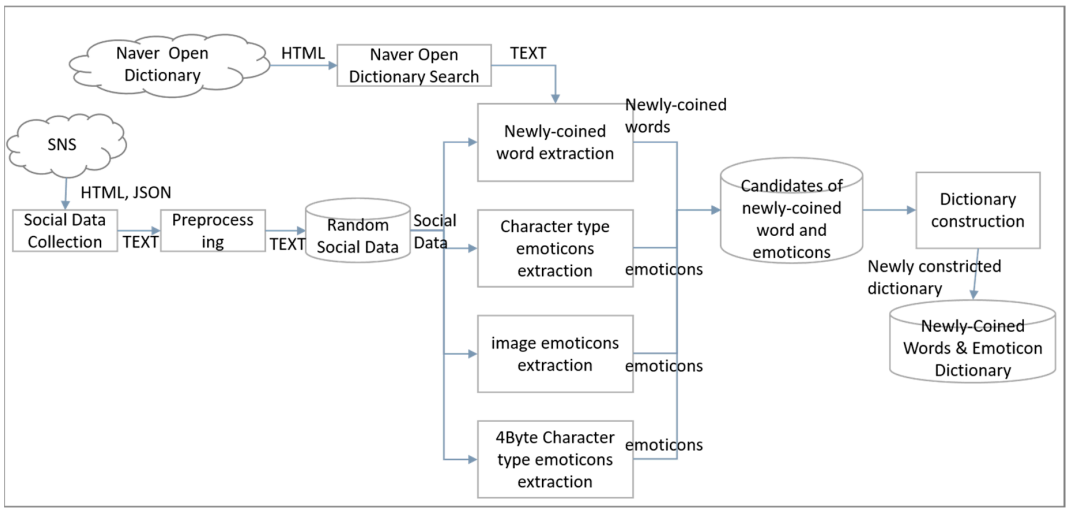

Figure 1. Newly coined words and emoticon extraction process.

3.1. Social Data Collection

We collect Twitter and Naver blogs to collect social data for extracting newly coined words and emoticons. Twitter is a social network service that communicates in short sentences. The characteristics of Twitter include ease of use, speed and scalability, and asymmetric networks. In terms of ease of use, Twitter is not burdened with writing short-sentence-oriented content with a 140-character limit. It also has an advantage that it can be easily linked with external web pages through mobile devices and APIs [1]. A blog can be a specific webpage format or a tool for creating it that can simply upload

articles (such as short mentions on daily activities or existing articles) [2]. Real-time Twitter tweets are randomly collected through Twitter search API. Twitter's search API provides about 1%–2% real-time tweets for all of Twitter's tweet. Twitter4j library developed on the java platform is used. Figure 2 is a function using the Twitter4j library. Because Naver blog does not provide its own search API, it develops and uses a crawler that collects Naver blog. The crawler periodically collects the contents of the Naver blog included in the "Entertainment/Art" category.

```
ConfigurationBuilder cb = new ConfigurationBuilder();
cb.setDebugEnabled(true)
.setOAuthConsumerKey("6yEys4QejSPkUQjn4TwBw")
.setOAuthConsumerSecret("f76AKMVfzc7r10dmuxR6eou7uKLJIOiYH9mpcyw3dG0")
.setOAuthAccessToken("206768244-
k85VKyGmxrFNM7DOU6yQr7GgCmVGbRCOeT2rboDV")
.setOAuthAccessTokenSecret("Qb3oHG19XnWIX3f2fll28RE2oPOkKjY2nJCxlrue4");
TwitterFactory tf = new TwitterFactory(cb.build());
Twitter twitter = tf.getInstance();
Query query = new Query("lang:ko");
result = twitter.search(query);
for(Status status : result.getTweets()){
  System.out.println("@" + status.getUser().getScreenName() + "\n");
  System.out.println(status.getId() + "\n");
  System.out.println(status.getText() + "\n");
  System.out.println(status.isRetweet() + "\n");
}
```

Figure 2. Collect tweets using the Twitter4j library [15].

3.2. Preprocessing Process

Randomly collected tweets are raw data, and preprocessing is required to minimize errors in the extraction of newly coined words and emoticons. Randomly collected tweets contain some unneeded information, such as usernames, URLs, hashtags, and so on. Unnecessary information is not included in the subject of pre-extraction and may be a factor in lowering the accuracy of emotional analysis. To remove unnecessary information, this study uses the regular expression pattern of the java platform. Figures 3–5 are functions used in the preprocessing process of the tweeter. Figure 3 is a function for removing a URL from a tweet, Figure 4 is a function for removing a user ID, and Figure 5 is a function for removing a hash tag. The URL, user ID, and hashtags from randomly collected tweets using URL removal function, user ID removal function, and hashtag removal function. Table 1 is an example of a tweet in which a preprocessing process of a Twitter is completed. In Naver blog, image type emoticons consist of HTML tags. Since you do not need to extract the image type emoticons, you should remove all content and tags except the <IMG> tag from your blog. Twitter and blogs that have been preprocessed are stored in the social data database for the dictionary extraction. As shown in Table 2, the dictionary extraction DB is composed of the original text ID number, the channel, and the content. The original number is a key value that is automatically generated when a Twitter and a blog are stored. A channel means a division between a Twitter and a blog, and a content means collected social data.

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

Figure 3. Uniform resource locator (URL) removal function [16].

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

Figure 4. User ID removal function [16].

```
public static String ClearTag(String content) {
  Pattern URL = Pattern.compile("(#[^ #\"]+)");
  Matcher m = URL.matcher(content);
  content = m.replaceAll("");
  return content;
}
```

Figure 5. Hashtag removal function.

Table 1. Example of Twitter pre-processing process.

Raw Source Data	Compensated Social Data
RT @parfaitfemi: And really so handsome that the movie is not the Olympics … #Pyeongchang Winter Olympics—Short Track 1500 m Men Finalists Sandor Liu (Hungary) https://t.co/LWPRXGyiBq	RT And really so handsome that the movie is not the Olympics … #Pyeongchang Winter Olympics—Short Track 1500 m Men Finalists Sandor Liu (Hungary)
RT @rifkrifk: Pyeongchang toilet since LOL LOL LOL men's toilet is left, cause the women are always right (right) https://t.co/TjoQGvI8iM	RT Pyeongchang toilet since LOL LOL LOL men's toilet is left, cause the women are always right (right)
RT @1theleft: Worldwide 5 billion people support the TV 2925 players in 92 countries The largest number in history # Pyeongchang Olympics # Peace Olympics # Daily Moon https://t.co/X8TPswXKWH	RT Worldwide 5 billion people support the TV 2925 players in 92 countries The largest number in history # Pyeongchang Olympics # Peace Olympics # Daily Moon

Table 2. Examples of social data for word extraction.

Id	Channel	Content
1	Twitter	RT @6882c93abe3b4a1: "Park Geun-hye, cares about jeong Yu Ra." said the horse riding system was a non-linear reality. The former vice chairman of Korea Racing Authority, Lee Sangyoung came out as a witness and said, "I have heard directly that former President Park Geun-hye chose to preserve his … "
2	Twitter	Discussions: 'Cooperation in the Left' by Park Geun-hye, 'Workers' Movement' Discussions: In the controversy surrounding the direction of the movement of withdrawal since the parliamentary cabinet was elected and Park Geun- · Labor Party · Labor Solidarity · Labor Front held a public debate on December 29th. The debate at the Korean Confederation of Trade Unions (KCTU) on the theme of 'struggle for the withdrawal of the Park Geun-hye regime and the worker's movement' was the purpose of advancing the movement and seeking joint activities of the left. … wspaper.org
3	Blog	<img src="https://storep-phinf.pstatic.net/ogq_56cdf4f9d81cd/original_11.png?type=p50_50" class="__se_img_el" alt="sticker image" width="185" height="160">
4	Blog	<img src="http://image.kmib.co.kr/online_image/2018/1122/612211110012860574_3.jpg" data-attachment-id="" alt="" class="se_mediaImage __se_img_el" id="SEDOC-1542871992468—1651403461_image_1_img">

3.3. Extraction of Newly Coined Words

We use Naver-open dictionaries as base data for extracting newly coined words from social data. The Naver Open Dictionary, a user-participation-open dictionary has 32 new languages including Korean, English, Chinese, and Japanese, registered on the website (opendict.naver.com). Users of open dictionaries can evaluate "good" or "dislike" for new words registered by other users and can register new ones in the open dictionary directly. The newly coined words are displayed on the page in the "real-time word" list as shown in Figure 6 and waits for evaluation by other users. In this study, we developed a dedicated web crawler to extract Naver-open dictionary. Figure 6 is the "real-time word" page of the Naver-open dictionary website that the web crawler will collect. The search conditions at the time of collection are limited to Korean. Since the Naver Open Dictionary is a user participation dictionary, it can be a dictionary word that does not match the dictionary meaning or is not frequently used in the SNS. To solve this problem, we store words and meanings of dictionaries containing more than 10 'likes' in the Naver Open Dictionary DB. We use the collected Naver-Open Dictionary to extract the newly coined words and emoticons from social data.

Figure 6. Naver Open Dictionary.

3.4. Separation of Word Unit

In the word segment separation step, the tokenizing operation for the word phrase is performed after separating the social data into the 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 a newly coined word and an emoticon is extracted by extracting only partial data of the entire social data. The boundary between the word and the word is decided based on the spacing. The words in spacing units in the original social data table for extracting newly coined words and emoticons are tokenized. Then, when we are performing emotional analysis, a raw social data table for emotional analysis including collected keywords is also tokenized in a space unit.

3.5. Emoticon Extraction

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 in Figure 7. Hangul expresses characters by combining the three elements of initial/neutral/longitudinal sound. Therefore, Hangul cannot express a character as a single element of initial/neutral/longitudinal sound. However, most of the emoticons consist of only the initial sound. Using these characteristics, a word consisting of only an initial sound in a tokenized word is judged as an emoticon and extracted. The extracted character type emoticons are stored in the emoticon candidate table. Second, we extract the emoticon of the image form as shown in Figure 8. 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 generally composed of <IMG> tags in HTML format. After extracting the <IMG> tag from the tokenized word, it finds unique

patterns of SNS companies only. For example, if there is a value of "sticker image" in alt attribute in the emoticon tag of image form in Figure 8, extract <IMG> tag. The extracted <IMG> tag is stored in the emoticon candidate table with a character type emoticon. Third, we extract the emoticons in 4-Byte Unicode character, as shown in Figure 9. 4-Byte Unicode characters have been extended from 2-Byte Unicode characters to 4 Byte, 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. of Japan. "Emoji" is supported by Apple and Google as well as SNS company Facebook. A 4-Byte Unicode character-type emoticon extraction process, such as "Emoji", examines all characters in the tokenized word to find a character encoded in 4-Byte Unicode. The extracted 4-Byte Unicode character type emoticons are stored in the emoticon candidate table as character type emoticons.

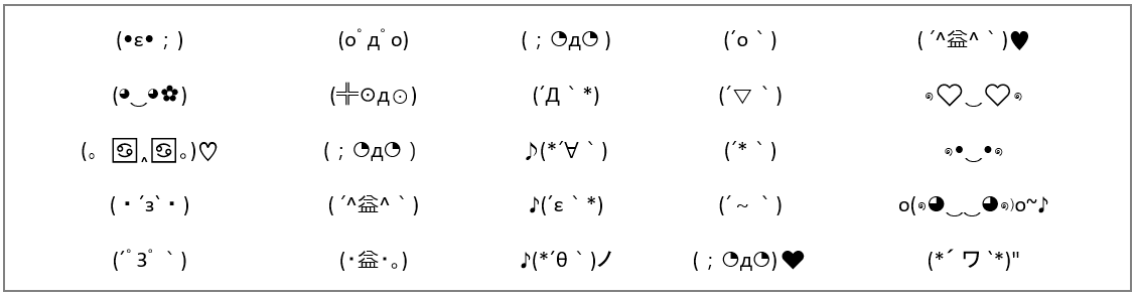

Figure 7. Character type emoticons [16].

Figure 8. Image type emoticons [16].

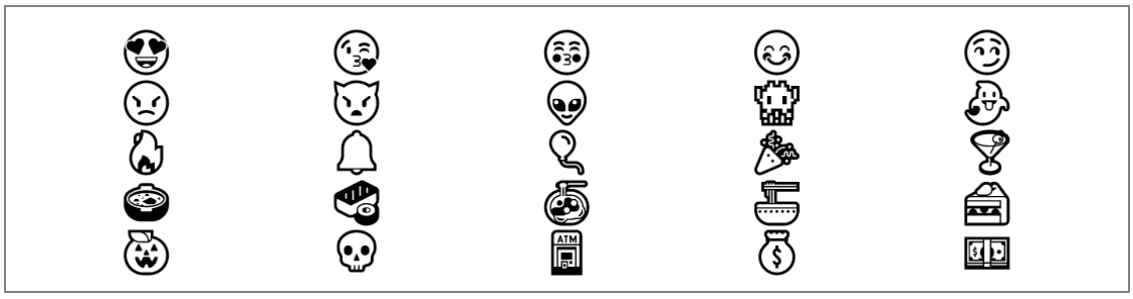

Figure 9. 4-Byte character type emoticons [16].

3.6. Dictionary Registration

If the extraction of the newly coined word and emoticon is completed, the registrant registers the newly coined word and emoticon emotion dictionary. The registrant categorizes the effective newly coined words and the effective emoticons in the newly coined word candidate table and emoticon candidate table. Valid newly coined words that are not included in the general emotional dictionary and the newly coined word dictionary include buzzwords or feed stuff items that represent political and social issues. The feed stuff is Korean slang, which is mainly used by teenagers. Among adults, the frequency use of feed stuffs is increasing, and the use of feed stuffs is also spreading rapidly in SNS. Valid emoticons that are not registered in the emoticons dictionary include character type emoticons,

Emoji emoticons, and image tag type emoticons. The classification of valid newly coined words and emoticons is done manually. To solve this problem, we developed a dedicated tool to clarify the classification. The dedicated tool can search the extracted newly coined words and emoticons in detail as shown in Figure 10 and can specify polarity and weight. Polarity can be classified into positive and negative, and weights can range from 1 to 5. The newly coined words and emoticons registered using the dedicated tool are stored in the emotion dictionary table.

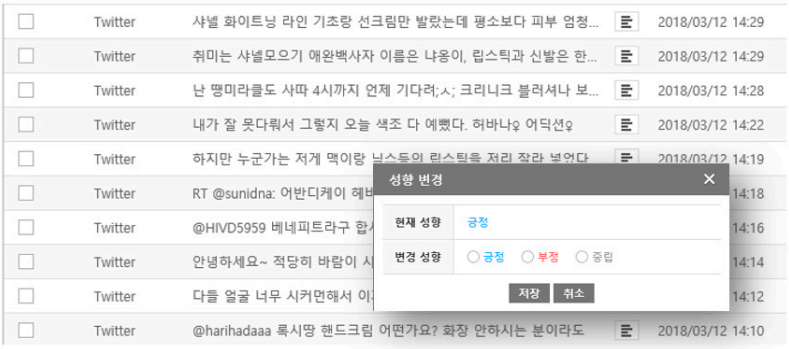

Figure 10. Classification of new words and emoticons using dedicated tools.

4. Emotional Analysis

Proposed emoticon and newly-coined words polarity decision system consists of enter process of collected keyword, social data collection process, contents pre-processing process, morpheme extraction process and emotional classification process in Figure 11.

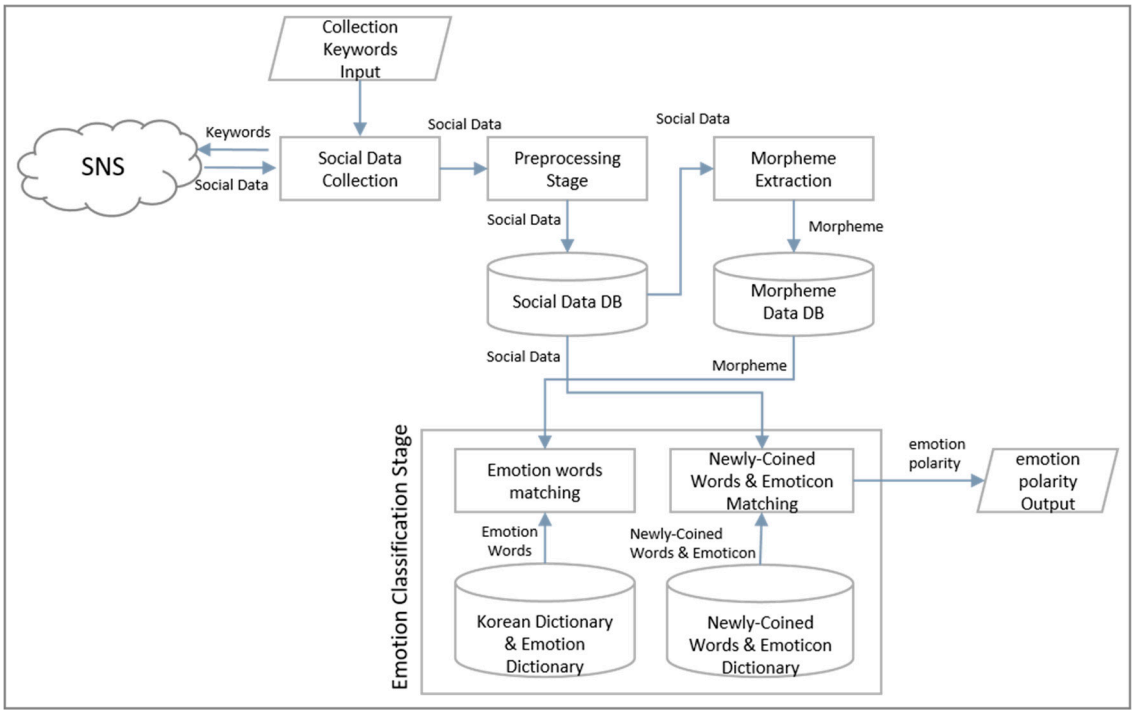

Figure 11. Emotional analysis process [16].

4.1. Entering Collected Keywords

Emotional analysis users select the topics to be subjected to emotional analysis for social big data emotional analysis and enter it. For example, in an emotional analysis of the social big data of the 2017

U-20 World Cup Korea vs. Portugal game held on 30 May 2017, the analysis users enter the search term as 2017.05.30, the search term is "U20" "Soccer", "World Cup", "Korea Portugal".

4.2. Collecting Social Data

When collecting social data, the collection process collects the collected data including the search word entered in the "input step of the collected keyword" using the API provided by the SNS vendor. The collected data containing the query are stored in the original message table. As shown in Table 1, the elements of the original message table are automatically generated keys as the original text number (ID), and the collected messages are the contents (CONTENT).

4.3. Preprocessing Process

The collected data containing the query are raw data, and a preprocessing process is required to minimize errors in emotion analysis. The collection data containing the query include some unnecessary information such as user name, URL, and hash tag. Unnecessary information is information that is not subject to analysis. Therefore, it may be a factor that lowers the accuracy of emotional analysis. To remove unnecessary information, this study uses the regular expression pattern of the java platform. Figure 4 is a function that removes the URL if the message contains a URL, and Figure 5 is a function that removes the user ID when the user ID is included in the message. Figure 6 is a function that removes a hash tag when a hash tag is included in the message. URL removal function, user ID removal function, and hashtag removal function are applied to the collected data including the search term, and the user name, URL, and hash tag entry are removed. The collected data, including the search term after the preprocessing process, are stored in the correction data table. The elements of the correction data table are the same as the original data table.

4.4. Morpheme Extraction

The morpheme is extracted by the morpheme extractor process in the message in which the preprocessing has been completed. In [9], morphemes is defined as the basic unit for analyzing language and the smallest grammar unit, that cannot be analyzed any more. Morphological analysis is also a process of separating morphemes from words and restoring them. In order to extract the morpheme of the message, this study uses the 2.4 version of the Comoran morpheme analyzer, Shin-Jun Su, provided in the form of a Java library. Unlike conventional morpheme analyzers, Comoran morpheme analyzer can analyze multiple phrases as a part of speech, allowing more accurate analysis of proper nouns including white space (movie title, restaurant name, song title, etc.). In this study, only the verbs, nouns, and adjectives that will match the emotional word among the parts of the morpheme extracted by the Comoran morpheme analyzer are used for emotional classification. The extracted morphemes are stored in the morpheme table. Elements of table consist of original number (ID) and morphology (MORPHEME).

4.5. Emotional Classification

In general, emotional classification of natural language is classified into analysis method based on emotion dictionary and analysis method through machine learning. The analysis method based on the emotion dictionary calculates how many times the emotion word is included in the natural language. In this paper, the emotion dictionary-based analysis method is used. In addition to the basic emotion dictionary, a new emulation dictionary and newly coined-word dictionary are separately constructed and used for emotion analysis. The basic emotion dictionary uses the Korean sentiment analysis corpus (KOSAC) of [10]. Korean Emotion Analysis Corpus (KOSAC) annotated 17,582 emotional expressions, using 332 newspaper articles and 7744 sentences as a commentary for constructing Korean emotion corpus essential for emotional analysis [10,17]. Elements of the emotion dictionary using the Korean emotion analysis corpus are composed of a dictionary code (SEQ), a dictionary name (WORD), a polarity (SENTIMENT), and a weight (INTENSITY). In Figure 12, the polarity of the emotion

dictionary is positive if the word is a positive word or negative if it is a negative word. The weights are weighted from 1 to 4, and the higher the number, the stronger the polarity. The newly coined words and emoticon dictionary are extracted by using the coined word and emoticon extraction process presented in this paper. In general, the analysis method based on emotion dictionary extracts morphemes from natural language through morphological analysis and whether the morpheme contains emotion words is computed. However, morphological extraction is impossible because the newly coined word and emoticon do not have a basic morpheme. In order to solve this problem, we first extract morphemes using morphological analyzer as well as an existing emotional analysis method and extract emotional words by comparing them with basic emotional dictionaries in extracted morpheme. In the second step, the newly coined word and emoticon are extracted from the completed data before the morphological analysis using the newly coined words and emoticon dictionary. Emotion words, newly coined words, and emoticons extracted from the message include polarity and weight, and polarity and weight are used for emotional classification. The emotion classification equation summarizes the weights among the words with positive polarity and adds all the weights among the words with negative polarity. Subsequently, the negative weight value is subtracted from the sum of the positive weight values, and the resulting value stores the corresponding value in the variable. Here, if the result value is more than 1, the sensitivity of the message is determined as affirmative. If the result is less than 1, the negative value is determined.

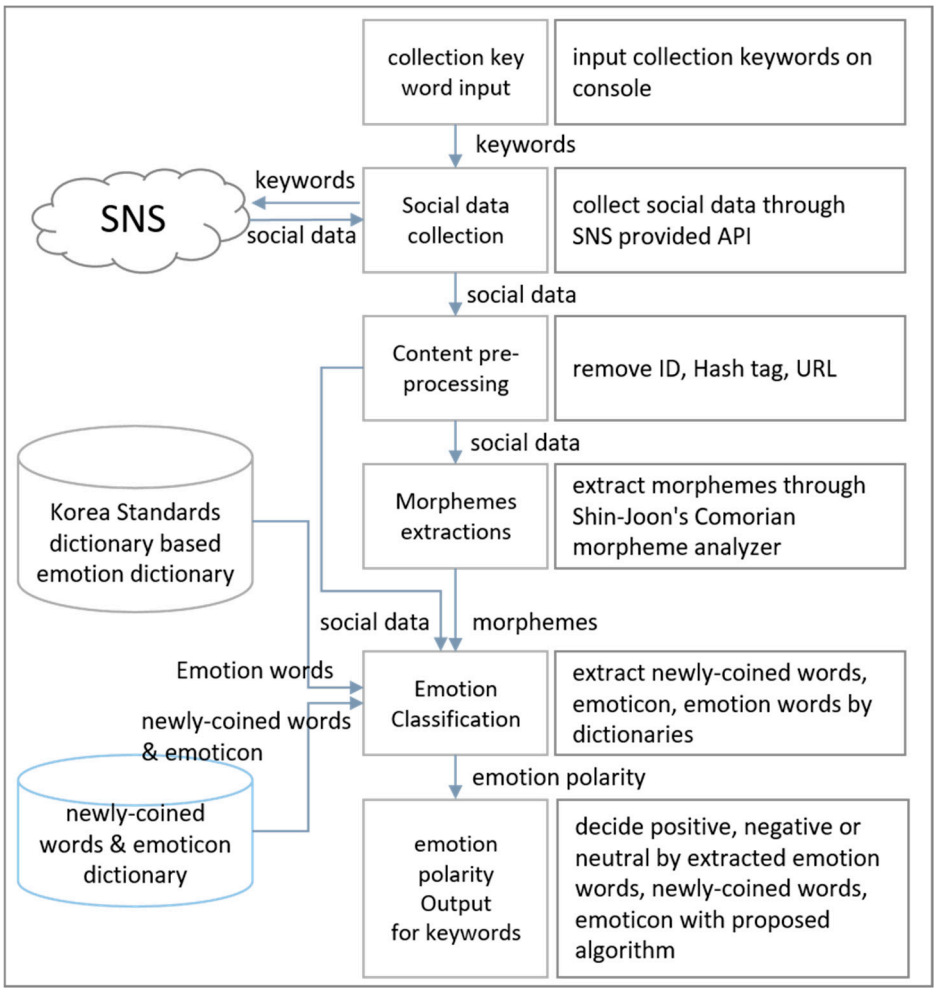

Figure 12. Emotional polarity decision algorithm [15].

5. Discussion

Based on newly coined words and emoticon extraction process presented in this paper, we have constructed a newly coined word and emoticon extraction and classification system and an emotional analysis system with newly coined words and emoticon based on the emotional analysis process in Section 4. The system is built on JAVA JDK 1.6 and CentOS 6.5. MYSQL 5.5 is used as the database to store social data. The social data are extracted only from Twitter. We used twitter4j library and the Java Wrapper of Twitter API, in order to randomly collect tweets in Korean. The collection period is from 7 March 2018 to 7 May 2018, and a total of 4,210,744 tweets were collected. The collected tweets were processed through a preprocessing process to remove the URL, user ID, and hashtag. We collected 90,373 Naver Open dictionaries using the crawlers developed in the JAVA language for extracting new words from the collected tweets. In the collected tweets, if the word in the open dictionary is included, it is judged as a newly coined word, and 55,996 cases were extracted. Also, 31,340 character-type emoticons and 1171 4-bit type character-type emoticons were extracted for the emoticons. Image-type emoticons are excluded because they are not provided by Twitter. On Twitter, we collected 725,928 tweets from 6 August 2018 to 7 September 2018, with the keyword 'idol.' As a criterion for measuring the accuracy of the emotion analysis method presented in this paper, a person directly classified tweets having a positive inclination and a tweet having a negative inclination in a tweet. The process of categorizing 725,928 tweets by hand requires a lot of time and labor, limiting the number of positive tweets to 926, negative tweets to 327, with a total of 1253 validation tweets. The evaluation items of this experiment were emotion analysis accuracy using existing Korean dictionary-based emotion dictionary, emotion analysis accuracy using new word and emoticon emotion dictionary, emotion analysis accuracy using existing Korean dictionary-based emotion dictionary, and newly coined word and emoticon emotion dictionar, respectively. The tweets are pre-verified and the URL, user ID, and hashtag of tweets are removed. The emotion dictionary based on the existing Korean dictionary uses Korean sentiment analysis corpus (KOSAC). Korean Sentiment Analysis Corpus (KOSAC) consists of morphemes. Therefore, we extract morphemes from the tweets for verification in emotional analysis using emotional dictionary based on existing Korean dictionary, and then extract emotional words by matching Korean sentiment analysis corpus (KOSAC). Shin Won-Su's Comorian morpheme analyzer 2.4 [10] was used for morphological extraction from the verification tweet. The measurement method verifies the correspondence between the tweet of positive inclination and the tweet of negative inclination classified by the person in the verification tweet.

Figure 13 shows the frequency of use of newly coined words and emoticons in tweets used in experiments. In total, 65.4% of the collected tweets had newly coined words and emoticons used in the experiment. SNS users usually use standard words, newly coined words, and emoticons in the tweets.

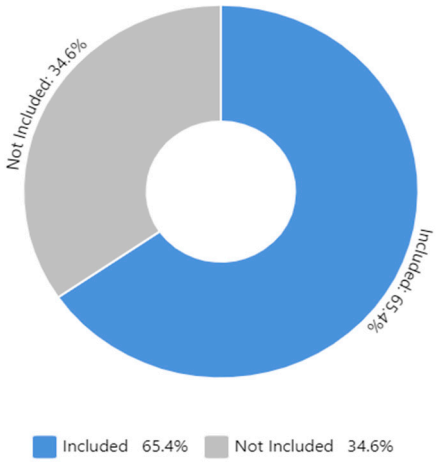

Figure 13. Frequency of newly coined words and emoticons used.

Figure 14 shows the results of the analysis on the positive tweets. The accuracy of emotional analysis by Korea standard word with emotional dictionaries, the accuracy of emotional analysis by the proposed newly coined words and emoticon dictionary, the accuracy of emotional analysis by Korea standard word with emotional dictionaries and the proposed newly coined words and emoticon dictionary are compared. The accuracy of emotional analysis by Korea standard word with emotional dictionaries is 53.5%. The accuracy of emotional analysis by the proposed newly coined words and emoticon dictionary is 56.3%. The accuracy of emotional analysis by Korea standard word with emotional dictionaries and the proposed newly coined words and emoticon dictionary is 80.2%. With Korean standard word with emotional dictionaries and the proposed newly coined words and emoticon dictionary, accuracy improvement of emotional analysis is 26.7%. The number of positive tweets is affected by events as like concert, donations, and showcases. For newly coined words and emoticons that are not included in the first analysis stage, polarity of positive tweets is decided only with Korean standard words with emotional dictionaries. After that, polarity of positive tweets with the proposed newly coined words and emoticon dictionary added to polarity of positive tweets are decided with Korea standard words with emotional dictionaries.

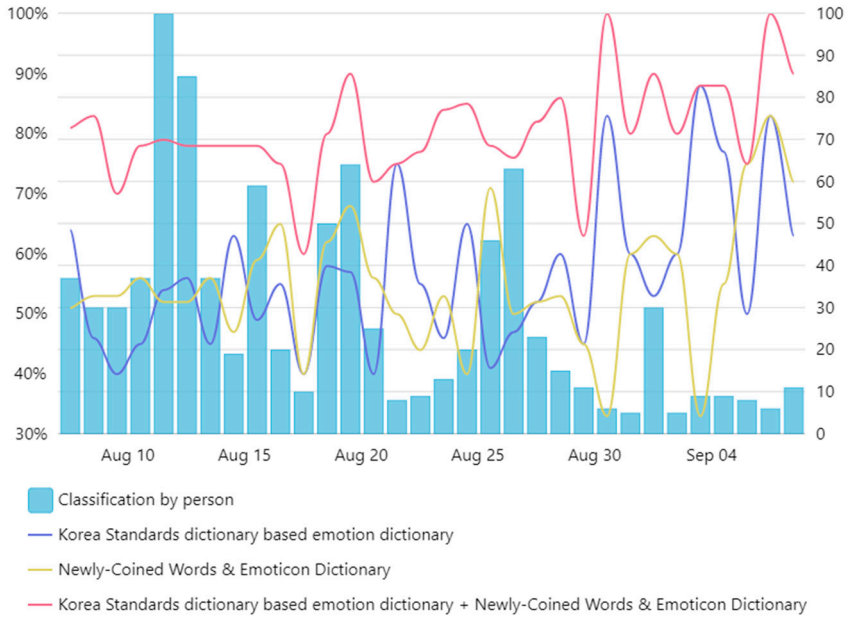

Figure 14. Accuracy analysis of positive tweets.

Figure 15 shows the negative tweet results. The average accuracy of emotional analysis using standard word based emotional dictionaries is 55.8%. The average accuracy of emotional analysis using new emulation and emoticon-based emotional dictionaries is 53.8%. The average accuracy of emotional analysis using standard word-based emotional dictionaries and new terms and emoticon-based emotional dictionaries is 80.1%, which is 24.3% higher than the existing emotion analysis. Emotional analysis results of negative tweets yielded similar experimental results as well as positive tweets. Emotional dictionaries based on the Korean standard word and the emotional dictionaries based on the newly coined word and emoticon showed a similar level of emotion analysis accuracy.

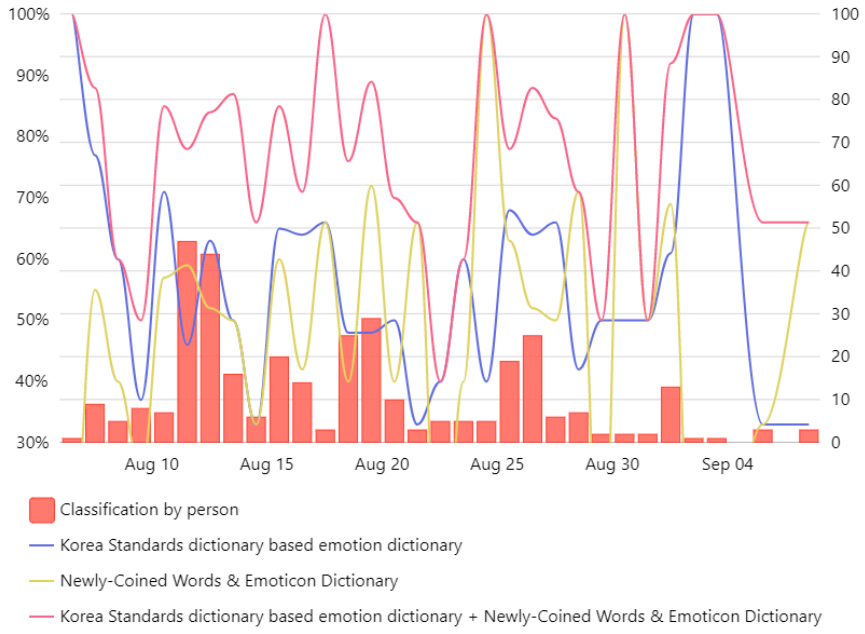

Figure 15. Accuracy analysis of negative tweets.

6. Conclusions

As mobile devices were developed and gained popularity, smartphones became a necessity in everyday life, and the exchanges of personal opinions about social issues or society happenings through smart phones enter into general use. As a result, the importance of analyzing individual opinions in SNS is increasing. Specifically, companies and governments have been interested in the results of emotional analysis on social data. However, research on emotion analysis techniques for newly coined words and various kinds of emoticons has been limited due to too quick modification and creation of newly coined words and emoticons.

In this research, a newly coined words and emoticons dictionary construction and analysis method is proposed. For newly coined words and emoticons dictionary construction, the morphological analyzer extracts the morphemes from the twit messages of Twitter, and the newly coined word and emoticon are extracted from the pre-processed data (morphological analysis results). This newly coined words and emoticons dictionary is used to decide the polarity of twit messages. In newly coined words and emoticons dictionary analysis method, the social data emotional polarity decision algorithm is proposed. The social data emotional polarity decision algorithm uses an emotion classification equation and the proposed newly coined words and emoticons dictionary. In order to evaluate the effectiveness of the newly coined words and emoticon in emotional analysis of SNS (twitter messages), we compared the results analyzed in the emotion dictionary based on the existing Korean dictionary and the analysis results based on the newly coined word and emoticon dictionary. After collecting about 700,000 twits that contain the keyword 'idol' from from 6 August 2018 to 7 September 2018 using Twitter's API, emotional analysis experiments showed that emotions, including new words and emoticon dictionaries, the accuracy of the analysis and the accuracy of the non-emotional analysis are, respectively, improved.

In order to fully understand the user base and to correctly analyze the relationship between the users and the options, user analyze can be added for further studies. This study has its limitations as it did not include detailed user information, since Twitter API does not provide it. Added user survey can provide more details about the relationship between the sentiment opinions and where and how the differences or correlations lie among certain users or certain group of users.

Author Contributions: Conceptualization, K.S.C.; Methodology, K.S.C.; Software, J.S.Y.; Validation, K.S.C. and M.-S.K.; Data Curation, J.S.Y.; Resources, J.S.Y.; Writing—Review & Editing, K.S.C. and M.-S.K.; Supervision, K.S.C.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kim, S.Y.; Park, S.T.; Kim, Y.K. Samsung-Apple patent war case analysis: Focus on the strategy to deal with patent litigation. *J. Digit. Converg.* **2015**, *13*, 117–125.
2. Bollen, J.; Mao, H.; Zeng, X. Twitter mood predicts the stock market. *J. Comput. Sci.* **2011**, *2*, 1–8. [CrossRef]
3. DiGrazia, J.; McKelvey, K.; Bollen, J.; Rojas, F. More tweets, more votes: Social media as a quantitative indicator of political behavior. *PLoS ONE* **2013**, *8*, e79449. [CrossRef] [PubMed]
4. Pak, A.; Paroubek, P. Twitter as a corpus for sentiment analysis and opinion mining. In Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, Valletta, Malta, 17–23 May 2010; Volume 10.
5. 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 (LSM 2011), Portland, OR, USA, 23 June 2011; pp. 30–38.
6. Go, A.; Bhayani, R.; Huang, L. *Twitter Sentiment Classification Using Distant Supervision*; CS224N Project Report; Stanford University: Stanford, CA, USA, 2009.
7. Kim, S.M.; Hovy, E. Determining the sentiment of opinions. In Proceedings of the 20th International Conference on Computational Linguistics, Geneva, Switzerland, 23–27 August 2004.
8. Yadollahi, A.; Shahraki, A.G.; Zaiane, O.R. Current State of Text Sentiment Analysis from Opinion to Emotion Mining. *ACM Comput. Surv.* **2017**, *50*, 25. [CrossRef]
9. Song, E. The Sensitivity Analysis for Customer Feedback on Social Media. *J. Korea Inst. Inf. Commun. Eng.* **2015**, *19*, 780–786. [CrossRef]
10. Kim, M.; Jang, H.; Jo, Y.; Shin, H. KOSAC: Korean Sentiment Analysis Corpus. In Proceedings of the Korea Information Science Society, 2013; pp. 650–652.
11. Oh, P.; Hwang, B. Real-time Spatial Recommendation System based on Sentiment Analysis of Twitter. *J. Soc. E-Bus. Stud.* **2016**, *21*, 15–28. [CrossRef]
12. Lee, G.H.; Lee, K.J. Sentiment Analysis of Twitter on Current Trends Extracted from Newspapers. *Korea Inf. Process Soc.* **2013**, *2*, 731–738.
13. Kim, K.; Lee, J. Sentiment Analysis of Twitter using Lexical Functional Information. In Proceedings of the KISS 2014 Conference, Busan, South Korea, 25–27 July 2014; pp. 734–736.
14. Hong, D.; Jeong, H.; Park, S.; Han, E.; Kim, H.; Yun, I. Study on the Methodology for Extracting Information from SNS Using a Sentiment Analysis. *J. Korea Inst. Intell. Transp. Syst.* **2017**, *16*, 141–155. [CrossRef]
15. Yang, J.S.; Chung, K.S. Newly-coined Words and Emoticon Polarity for Social Emotional Opinion Decision. In Proceedings of the 2019 IEEE 2nd International Conference on Information and Computer Technologies, Kahului, HI, USA, 14–17 March 2019; pp. 126–130.
16. Yang, J.S.; Chung, K.S. Newly-Coined Words and Emoticons Dictionary Construction for Social Data Sentiment Analysis. In Proceedings of the 2019 11th International Conference on Future Computer and Communication, Rangoon, Myanmar, 27 February–1 March 2019; pp. 126–130.
17. Jang, K.; Park, S.; Kim, W. Automatic Construction of a Negative/positive Corpus and Emotional Classification using the Internet Emotional Sign. *J. KIISE* **2015**, *42*, 512–521. [CrossRef]