DOI:10.6129/CJP.201909_61(3).0003
中華心理學刊 民 108，61 卷，3 期，209-230
Chinese Journal of Psychology 2019, Vol.61, No.3, 209-230

A Text Analysis Approach to Analyzing Gender Differences in Breakup Posts on Social Media

Lee-Xieng Yang[1, 2] and Ching-Fan Sheu[3]
Department of Psychology, National Chengchi University[1]
Research Center for Mind, Brain, and Learning, National Chengchi University[2]
Institute of Education, National Cheng-Kung University[3]

Two approaches to text analysis have been applied in the current work to investigate gender differences in breakup posts on social media in Taiwan. First, we calculated the probabilities of the types of words, such as personal pronoun and other word categories based on the Chinese Linguistic Inquiry and Word Count(LIWC), occurring in the posts to predict author's gender. The results showed that personal pronoun outperformed other word types at predicting gender for social media break-up posts. Second, we conducted stylometric analysis on these posts to extract keywords for different gender. The occurring probabilities of these keywords were then used to predict the author's gender of the post. The results showed that including keywords in the top one percent as predictors enabled a model to perform better than the first approach. A network analysis was carried out, respectively, for each gender to examine the psychological and linguistic features of these keywords and their relationship with reference to the Chinese LIWC. The typical features, defined in terms of the centrality indices, such as word types of verb, adverb, relative, social process, biological process and cognitive mechanism were found to be common for both gender. However, features of affection words, sexual words, and negate words showed up only for breakup posts authored by females. We conclude that among Taiwanese users of social media females were more likely than males to make affective statements.

Keywords: *gender differences, social media, stylometric analysis, text analysis*

Background

The emergence of the concept of big data has begun to shift the focus of psychological research from how to properly collect human data to how to properly retrieve and analyze already generated data from cloud storage (Serfass, Nowak, & Sherman, 2017). This change has been made possible by machine learning techniques that allow Web crawlers to automatically transfer webpage text into mathematical matrices, from which features can be extracted by text analysis. As this research approach is still new to psychologists, it is necessary to explore and examine the extent to which it can benefit psychological research. Using breakup posts on social media as the target, we thus demonstrate how to combine text analysis techniques in machine learning and psychology to extract gender markers and their psychological meanings.

Recent studies have shown that text features can be used to predict an author's personal attributes, such as race, age, gender, and personality (Kern, Eichstaedt, Schwartz, Dziurzynski et al., 2014; Park et al., 2014). The features normally consist of keywords, phrases, or n-grams that best represent the sample text(s). As these features emerge from the text analysis rather than being predetermined, this approach is also known as open vocabulary analysis (Schwartz et al., 2013).

Parallel to engineers in computer science, the psychologist Pennebaker and his colleagues developed their own text analysis in an attempt to provide the psychological profile of the author according to the catalog of words s/he uses in their LIWC dictionary (Pennebaker, Booth, & Francis, 2007). In the LIWC

dictionary, common English words are sorted out in 71 categories according to their linguistic and psychological features, such as pronoun, noun, verb, cognition, positive emotion, negative emotion, and so on. Thus, the psychological status of the author can be revealed by the distribution of the predefined LIWC categories of the words s/he uses. As these categories are fixed and predefined, this approach is also known as closed vocabulary analysis (Schwartz et al., 2013).

Although a past study showed that open vocabulary analysis outperforms closed vocabulary analysis (Schwartz et al., 2013), we consider this conclusion to be unfair because the two analyses were actually developed for different purposes. Open vocabulary analysis is simply used to extract the features of texts, with no reference to psychology. In contrast, closed vocabulary analysis is used to delineate the psychological states behind words. Thus, we sought to combine these two analyses for use in psychological studies. To this end, and to demonstrate how to conduct an Internet data-based study in psychology, we applied these two analyses to analyze relationship breakup posts scraped from social media sites in Taiwan in an attempt to show the gender differences in the posts. This study presents many innovative research methods, including how to perform Chinese word segmentation, how to combine open and closed vocabulary analyses, and how to visualize psychological attributes in terms of the LIWC dictionary.

Method

Data Source and Apparatus

The data comprised 1,311 posts with "breakup" as the tag, scraped from the relationship forum on the largest anonymous social media site, Dcard (https://www.dcard.tw/f). On Dcard, the author's information, such as name, id, age, and location, is always hidden, except for gender. The ratio of male to female registered Dcard users is approximately 4:6.

Procedure

Similar to the procedure used by Kern, et al. (2016), this study consisted of three stages: data-collection, data-preprocessing, and data analysis (see Figure 1). The processes for all stages were programmed in R. In the data-collection stage, we used our own script to call the Dcard application programming interface and collect posts tagged as "breakup" from February 1 to June 8 2017. This resulted in 1,311 posts. In the data-preprocessing stage, we cleaned the downloaded texts by removing nonsense symbols and html code, and by performing Chinese word segmentation. Unlike English, there is no space between words in Chinese sentences. Thus, it is impossible to identify all individual words by checking the spaces surrounding them. We broke down every Chinese sentence into single words using the word-segmentation algorithm in jieba, which is an open source project for Chinese word segmentation that can be easily accessed by calling the package {jiebaR} in R. The basic principle of the jieba algorithm is to identify a valid Chinese word by checking whether there is a match for it in the corpus. The downloaded posts became string vectors composed of single Chinese words through the data-preprocessing stage. Subsequently, two types of text analyses were conducted with the primary goal of comparing their usefulness in extracting the features of texts, and the secondary goal of understanding the differences between the psychological attributes revealed by the breakup posts of male and female authors.

Results

Demographic Analysis

Among the 1,311 posts, approximately 29% (383/1311) were written by males and 71% (928/1311) by females. This suggests that female Dcard users are more willing than the males to share their breakup stories. However, there was no significant difference in the number of words between the male and female posts (mean words per post = 255.53 for males and 257.29 for females), $t(1309) = 0.10$, $p = .92$. The number of comments received also did not differ between genders (i.e., mean number of comments = 18.18 for males and 18.98 for females), $t(1309) = 0.23$, $p = .82$. These results show that if a gender difference exists, it is not evident in

the surface structure of posts.

Closed Vocabulary Analysis

As the closed vocabulary analysis delineated the psychological profiles of authors in terms of the distributions across the word categories chosen before the study, we first computed the probability of occurrence of each of 10 Chinese personal pronouns (i.e., 你 , 我 , 他 , 妳 , 她 , 你們 , 我們 , and 他們) used in the posts, according to the formula $p(ppronoun|post) = \frac{freq(ppronoun, post)}{lengt|(post)}$, where *freq*(*ppronoun, post*) is the frequency of a particular personal pronounin a post and *lengt*|(*post*) is the number of words in the post. Thus, for each post, we had a 10-element vector representing the distribution over the probabilities of the 10 pronouns. Subsequently, a logistic regression model with the 10 probabilities as the predictors was fit to the gender data (1: female, 0: male). The accuracy of the model was .85 and the goodness-of-fit measured by the Akaike information criterion (*AIC)* was 1067, *df* = 1300. Although the prediction accuracy was good, the use of pronouns may reflect a general gender difference rather than a difference specific to relationship breakup.

Therefore, we tested the performance of the logistic regression model with 12 LIWC word categories as the predictors. These categories were considered to be relevant to breakups, including words related to cognition, positive emotion, negative emotion, anxiety, sadness, and so on. The model did not perform well, as the prediction accuracy was .71 (*AIC* = 1559.2, *df* = 1298). Even worse, the model performance did not increase when the number of predictors increased to 22 (i.e., 12 LIWC word categories and 10 pronouns) [accuracy = .72, *AIC* = 1569.4, *df* = 1288], or when all 70 LIWC word categories were included as the predictors [accuracy = .72, *AIC* = 1600.6, *df* = 1240]. These results imply that the LIWC word categories are less sensitive than the personal pronouns in detecting the gender of the author of a breakup post.

Open Vocabulary Analysis

Because we sought to detect the keywords that could best distinguish between the two genders, stylometric analysis was chosen in this study. Although the TF-IDF algorithm is well-known for extracting the keywords of texts, we chose the Zeta test to conduct the open vocabulary analysis. We preferred the Zeta test because it can find the keywords (or short phrases or n-grams) that can best distinguish the works of two authors. These keywords are the words most frequently used by one author and least frequently used by the other.

To implement stylometric analysis, there must be at most two authors and at least two works for each author. Therefore, we randomly aggregated the posts as two "works" (of 29,223 and 29,175 words) for males and three "works" (of 45,346, 48,534, and 49,836 words) for females. Subsequently, the tokens for the works of each gender were generated by randomly replacing the words in the works of each gender. As a result, there were 18 male tokens and 47 female tokens, each of which was about 3,000 words. The Zeta test identified 811 keywords preferred by males and 695 keywords preferred by females. We show the first 50 male and female keywords, respectively, in the word clouds in Figure 2.

For a comparison with the closed vocabulary analysis, we also built up the logistic regression model to predict the genders of the authors with the probabilities of the keywords extracted by the Zeta test. Of course, the more keywords included as predictors, the better the regression model can perform. Thus, we gradually increased the proportion of keywords used in the logistic regression model from the first 1% to the first 19% of the keywords for both genders. Using the first 1% of keywords (8 for males and 6 for females), the prediction accuracy was .71, *AIC* = 1536.3, *df* = 1296, which was just as poor as the model with all 70 LIWC word categories as the predictors. However, the prediction accuracy improved, eventually reaching .90 (higher than the model with personal pronouns as predictors), as the proportion of keywords increased to 19% (see Figure 3). Thus, overall, the open vocabulary analysis performed better than the closed vocabulary analysis in detecting the keywords that distinguish between the two genders.

Psychological Reality Revealed in Breakup Posts

Although the open vocabulary analysis was evidently better than the closed vocabulary analysis for extracting the keywords from the posts, it still did not tell us which psychological attributes these keywords referred to. This is a limitation of open vocabulary analysis, but a benefit of closed vocabulary analysis. Because the word categories in the LIWC dictionary are intended to correspond to psychological-linguistic attributes, why not check the psychological attributes of the keywords extracted by the Zeta test to understand the gender differences in the breakup posts?

To this end, we conducted a network analysis using the LIWC categories corresponding to the first 19% of the keywords (154 male and 132 female words), as the model with these keywords as the predictors had a high level of accuracy (i.e., .90). According to the LIWC dictionary, most of the words could be classified under multiple categories. That is, when a word is read, the psychological feelings corresponding to the categories corresponding to that word occur together. Therefore, a network was established with the LIWC categories as vertices and the co-occurrence frequency between these categories as edges. In this network, the more central a category is, the more frequently the psychological attribute corresponding to it is mentioned (see Figure 5). A comparison of the two panels indicates that the breakup posts of both genders tended to use words referring to social mechanisms, cognitive mechanisms, and relationships, as well as biological terms. However, only females used words referring to affection, sex, and negation, as well as pronouns. Thus, it is suggested that females express more sentiments than males in their breakup posts.

Conclusion

Open vocabulary analysis is suitable for extracting the keywords of texts. However, the psychological meaning referred to by the keywords needs to be identified using the LIWC dictionary. Thus, a hybrid approach for data analysis is suggested for future psychological studies.