SSRG International Journal of Economics and Management Studies
ISSN: 2393 – 9125 / https://doi.org/10.14445/23939125/IJEMS-V7I7P102

Volume 7 Issue 7, 5-9, July, 2020

Original Article

Predicting Neologisms for Marketing:A Text Mining Approach

Sang-Uk Jung[1], Jungho Byun[2], Seongyeol Bae[3], Donghwi Song[4]

[1,2,3,4] *Business School, Hankuk University of Foreign Studies, 107 Imun-ro, Dongdaemun-gu, Seoul, Korea, 02450*

Received Date: 26 May 2020
Revised Date: 03 July 2020
Accepted Date: 07 July 2020

Abstract *- An increasing number of companies rely on neologisms when implementing their marketing strategy. However, companies recognize that indiscriminate use of neologisms can haveabadimpactonmarketing, requiring them to evaluate the pros and cons of the neologism they apply. To help on these issues, this research focuses on creating a model predicting neologisms that are appropriate for marketing use. Using data collected with web-crawling from Korea's largest community website, 415,000 terms for 6 forums are examined with network analysis, text mining, and logistic regression. We find that 'Negative', 'Summary', 'Korean' are the most meaningful variables when predicting appropriate neologisms for marketing use in Korea. Our model predicts that the 'Ja-gang-du-chun' will be a buzzword next year. Up-to-date results will come out with the updated and supplementary data sets. These findings suggest a way for the practitioner to predict a buzzword and how to use it in marketing.*

Keywords *- Neologism, Marketing Intelligence, Online Community, Text Mining, Korean.*

I. INTRODUCTION

While many companies today recognize the increasing commercial value and importance of neologisms in marketing, related research has been scarcely done. Ding Ying(2008) finds that neologisms are created by events that have a big influence on people, such as political events, economic events, and the introduction of new culture or technology. The newly made neologisms are used by people mainly for two psychological reasons, the wants of people to express themselves as different individuals compared to others, and the wants of people to keep up and not to be left behind with the majority of the society.

While Ding Ying(2008) examines the creation and classification of neologisms and the spread and use among the general public, this article focuses on what are the main factors that produce a neologism from a linguistic perspective. This research aims to develop the model that predicts Korean neologisms, usingweb-crawled large-scale word corpus data from one of the largest Korean community websites, Dcinside. A large-scale web-crawled data use offers some advantages of presenting each phase of neologism's life cycle – occurrence, growth, decline, extinction, andrebirth – with detailed examples (Kilgarriff & Grefenstette, 2003).

The remainder of this article is organized as follows. In the next section, we provide the details of the collection and preprocessing of data and describe the methodology used for current research. Then we discuss the empirical findings from our analysis. We conclude with the managerial implication of our findings, limitations of this study, and potential future study.

II. BACKGROUND

While previous research onneologisms is mostly about the detection of the coinage of new words and change in part-of-speech (POS), study about Korean neologisms are scarcely done.

There are two ways to discover new word forms. The first is to filter the unknown words by creating an exclusion list of known words. Filtering using the exclusion list is the most common way so far, and various methods such as Romary et al. (2004) andOllinger and Valette(2010) have been suggested. The downside of this method is that it is based on a simple heuristic, so it has to be validated by experts in the end.

The second is to apply various statistical methods such as logistic regression and machine learning to the historical corpus to detect neologisms (Wang and Wu, 2017).By using historical corpus, we can find out when a new word appeared. The cumulative distribution of new words occurrence increases exponentially, which can be recognized as neologisms bypassing certain thresholds.

While the reasons for adopting new words are varied, there are factors that predict where new words emerge as neologisms. Previous research has focused on finding factors as reasons affecting the appearance of neologisms. These factors include the diversity of linguistic contexts (Stewart and Eisenstein, 2018),demographics (Eisenstein et al., 2014), geography (Stewart and Eisenstein, 2018), the number of occurrences of the word in the corpus and the distribution of its occurrence over time. By using various factors from existing research and machine learning's various classification techniques, we can predict whether or not new words will be accepted as neologisms (Written and Frank, 2005).

Current research attempts to combine the streams of the two important research methods mentioned above. We use exclusion lists as a filter and try to predict which word would be accepted as neologisms using one of the classification methods of machine learning called logistic regression.

III. EXPERIMENTAL SETTING

A. Datasets

Our selection of community websites for the neologism analysis was based on the criteria of diversity, activeness, anonymity, and openness of the website. Inaccordancewiththeseselectioncriteria,weselectedDcinside whichis Korea's largest community website with more than 12,000 forums, each with its own theme such as a game, entertainment, sports, education, travel, etc. Because approximately 820,000 posts are generated, and large data packets are consumed every day, it provides a favorable environment in which neologisms are created and easily spread to other media.

Table 1. Centrality Measures of Top 5 Forums

Name of Forums	Degree	Betweenness	Closeness	Eigenvector	Page Rank
MCG	9	1.333	0.200	0.928	0.191
KDG	9	2.000	0.167	0.715	0.337
KBG	7	0.667	0.167	0.322	0.052
ISG	7	0.333	0.167	0.181	0.096
DFG	5	0.000	0.122	0.106	0.055

Due to the extensive size of the dataset, we further limited our data source to the most popular forums. Tomeasurethepopularityofeachforum,weusedthemonthlyav eragescoreofcentralityinthenetwork betweenJanuary 2017 and June 2019. Various centrality measures such as degree, betweenness, closeness, eigenvector, and page rank are based on the assumption that having a specific position in the network has a greater impact on other points (Newman, 2010). That is, neologisms used in forums with high centrality are likely to spread to or have a greater impact on neologisms in other forums.

To measure the centrality of forums, when other forums mentioned specific forums, the network was considered connected, and the direction was taken into account. Because of the variety of words and terms used to refer to the forum, we considered this in our research. For example, Korean Baseball Gallery, which is the official term for the gallery of Korean baseball, is often called in shorter terms such as KG', or KBG. Information on the top five forums selected by various centrality measures is as shown below in Table1. In the table, the name of each gallery is abbreviated for better readability.

If a particular post gets a lot of likes or hits, the post goes to the hot category and appears at the top of the website. For this study,we collect the scraped data from postsinthehotcategoryby top 2 forums, MCG and KDGbasedontheassumptionthat threads with high likes and views have a high possibility of containing words that people think are trending.

Among the trending category threadsfrom April 30, 2019,to June 6, 2019, a total of 2,065 threads were randomly selected,especially 1,125 threads from Korean Baseball Gallery and 940 threads from the Other TV Program Gallery.The thread included not only the main contents of the thread but also the comments that averaged over 200 per thread, sowedecideditwas enough data to conduct textualdata analysis. This turned out to be true when we converted text data to corpuses and created a Term Document Matrix(TDM).Itcreated more than a 2.6million terms in the Korean Baseball Gallery and 1.4 million terms in the Other TV Program Gallery.

B. Neologism Selection

One important aspect of finding neologisms is the appropriate specification of which word would be considered as neologisms. To ensure the consistency of our results with existing studies, we applied thefollowing three screening criteria to restrict the words we analyzed.

First, because less-frequent terms do not match the purpose of the study, sparse data was removed before text analysis. However, to maintain the diversity of terms and the volume of the data, were moved all terms in the corpus whose sparsity is greater than 0.9999,which ended up with 45,000 terms from Korean Baseball Gallery and 25,000 terms from Other TV Program Gallery.

Second, the two Term Document Matrix were combined into one, and the frequencies of the terms were added together. As expected, words used for grammar purposes were more frequent than neologisms. In the process of cleansing word data, we tried to definea word with the same definition as one, but there were variables taking into account the nuances of the word used in the regression analysis. Therefore, other words with the same definition were combined with one word under conditions with similar nuances. In addition, it was necessary to interpret the context in which words were used in the purification process, which was done by the subjectivity of theresearchers.

Third, words rarely covered in mass media have been removed. Because mass media is one of the main paths in which neologisms are accepted, the fact that it was rarely covered in mass media means that the probability of being accepted as neologisms is very low.These selection criteria result in 242 candidates of neologisms.

C. Features

We selected 16 features and explored the effect on accepting neologisms. All of these features are dummy variables. The first selected features are two sociolinguistic characteristics of languages, especially the Korean language. Since most dialectal neologisms seem to have a negative meaning (Liu et al., 2013), we conjecture that new words with negative meaningsare more likely to be accepted as neologisms. *Negative*indicates whether the word has a negative meaning. A neologism often starts with slang or jargon in a particular area and generally meets the needs created by new technologies or new social environments. *Specific*indicates whether the word is used in a particular area.

The second selected features are fourteen grammatical characteristics of Korean languages. This includeswhether there is a linking sound or liaison (*Continual*) or consonant assimilation (*Similar*) or phonological addition (*Add*) or phonological deletion (*Deletion*)orabbreviations (*Summary*) or acronyms (*Extract*) or loanword (*Foreign*) or new word (*New*) or dialect (*Dialect*) or prefix (*Prefix*), suffix (*Suffix*), pure Korean (*Korean*) or it is used independently (*Independent*), or the word is spoken in a real conversation (*Spoken*).

D. Research Method

$$logit(Y) = ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta_1 Negative_1 + \beta_2 Specific_2 + \beta_3 Continual_3 + \beta_4 Similar_4 + \beta_5 Add_5 + \beta_6 Deletion_6 + \beta_7 Summary_7 + \beta_8 Extract_8 + \beta_9 Foreign_9 + \beta_{10} New_{10} + \beta_{11} Dialect_{11} + \beta_{12} Prefix_{12} + \beta_{13} Suffix_{13} + \beta_{14} Independent_{14} + \beta_{15} Korean_{15} + \beta_{16} Spoken_{16} \quad (1)$$

where π is the probability of the event, α is the Y-intercept, βs are regression coefficients, and *X*s are a set of predictors.

To test the effect of various features we selected on the acceptance of neologisms, we fitted the logistic regression model above in (1) with our selected data. α and βs are estimated by the maximum likelihood (ML) method(Ahmed and Ahmed, 2019).

IV. RESULTS

A. Empirical Results

The results of model (1) are reported in Table2below. According to the model, the log of the odds of a new word being accepted as a neologism was negatively related to a word with negative meanings ($p < 0.001$), was negatively related to a word made of abbreviation ($p < 0.001$), was negatively related to a word of pure Korean ($p < 0.001$),

Table 2. Results of Logistic Regression Analysis

Coefficient	Estimate β	Error
Negative	-3.562^{***}	0.532
Specific	0.181	0.449
Continual	0.614	0.890
Similar	0.560	0.670
Add	1.282	1.103
Deletion	2.354^{*}	1.397
Summary	-1.818^{***}	0.519
Extract	0.100	0.497
Foreign	0.212	0.655
New	-0.725	0.552
Dialect	-0.741	0.817
Prefix	0.578	1.140
Suffix	0.228	1.103
Independent	0.417	0.870
Korean	-2.215^{***}	0.498
Spoken	1.343^{*}	0.655

*Significant codes: 0'***' 0.001'**' 0.01'*' 0.05'.' 0.1' '*

We run stepwise regression to select a reduced number of predictor variables resulting in a final model. The function chose a final model in which other variables except for*Negative, Summary,* and *Korean* are removed from the original full model. To choose the best model that has the lowest classification error rate in prediction, we compare the performance of the full and the stepwise logistic models. The prediction accuracy of the full and stepwise models is as follows, respectively: 0.408 and 0.395, which shows that the performance of the reduced model is similar to the full model. Because the reduced model decreases the complexity of the model without compromising its accuracy, we select the simpler reduced model as our final model.

Table 3. Results of Logistic Regression Analysis

Coefficient	Estimate β	Error
Negative	-3.562^{***}	0.532
Specific	0.181	0.449
Continual	0.614	0.890

*Significant codes: 0'***' 0.001'**' 0.01'*' 0.05'.' 0.1' '*

Results of our final model shown in Table 3 suggest that negative word has less likelihood of spread to mass media, for it has been firmly accepted as undesirable. Abbreviation or reduced word has less likelihood of spread to mass media for it has typical awareness of subcultural factor appears on the internet. A Word that is composed of only pure Korean words has less likelihood of spread to mass media, for it encounters common use of loanwords in the globalized society.

The results in Table 3 give an important managerial implication to marketing practitioners. For a successful marketing campaign, it is better to rely on the use of a language that expresses positive thinking and original expressions and uses a loanword rather than pure Korean.

B. Validation of the Results

We conducted Calibration work to re-verify the model. The process is to applicate past neologisms, which are given on 'Monthly Dcinside' – a period covering January 2017 through December 2018 - into the model and measure each word's value. Table 4 below shows the result, and we confirmed validation of the model by the point that a high level of prediction value is presented in general.

C. Prediction of Neologisms

Based on the obtained model, our approach to predict the possibility of spread – a wide range reaches mass media – targeted neologisms which are on growth phase in the life cycle. The neologisms are adopted from 'Monthly Dcinside' – a period covering January 2019 through December 2019 – official source as same as used in calibration work.

We applied function 'predict' to the final model, and Table 5 shows the result of each neologisms' value which indicates predictability. According to this, '자강두천', which has the closest value to 1, takes the highest level of utility in mass media.

V. CONCLUSION

To prove the result of the study, we have observed the mass media to spot the term 자강두천 in usage and have successfully accounted for numerous utilizations of the term in various fields. A politician used the term 자강두천 in his political statement, the various major press has used the term for the title and contents of their articles, and thumbnails of new media utilized the word countlessly. Especially, thumbnail, which is an image that explicitly shows the viewer or reader about the content of a media, having marketing effect of attracting viewers have used the term heavily.

The results of the prediction showed a positive accuracy rate considering the fact that this research was conducted on a single data source. Further researches regarding more data sources and language characteristics as variables should improve the model, resulting in better guidelines for actors of business who desire to use neologisms for marketing.

ACKNOWLEDGMENT

This research was supported by Hankuk University of Foreign Studies Research Fund.

REFERENCES

[1] Y. Ding, The study on the Korean vogue words, thesis, Chonnam National University, Kwangju, Korea,(2008).

[2] A. Kilgarriff, and G. Grefenstette, Introduction to the special issue on the web as corpus, Computational linguistics, 29 (2003) 333–347.

[3] M. E. Newman, Networks,An Introduction, Oxford, England: Oxford University Press, (2010).

[4] L. Romary, S. Salmon-Alt, and G. Francopoulo.Standards are going concrete: from LMF to Morphalou, In WorkshopEnhancing and Using Electronic Dictionaries, Geneva, Switzerland.

[5] S. Ollinger and M. Valette, La créativitélexicale:despratiquessociales aux textes. In Actes del I CongrésInternacional de Neologia de les llengüesromaniques, volume Publicacions de l'InstitutUniversitari de LingüísticaAplicada (IULA) de laUniversitatPompeuFabra (UPF), Barcelona, Spain, (2010) 965–876.

[6] I. Witten and E. Frank, Data Mining: Practical machine learning tools and techniques, 2nd Edition, Morgan Kaufmann, San Francisco, (2005).

[7] K. Wang and H. Wu. Research on neologism detection in entity attribute knowledge acquisition. In 2017 5th International Conference on Machinery, Materials and Computing Technology (ICMMCT 2017). Atlantis Press, (2017).

[8] I. Stewart and J. Eisenstein. Making fetch happen: The influence of social and linguistic context on nonstandard word growth and decline. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, (2018) 4360–4370.

[9] J. Eisenstein, B. O'Connor, N. A Smith, and E. P. Xing, Diffusion of lexical change in social media,PloS one, vol. 9(11) (2014) e113114.

[10] T. J.Liu, S. K.Hsieh, andL. Prevot, Observing features of PTT neologisms: A corpus-driven study with N-gram model. In Proceedings of the 25th Conference on Computational Linguistics and Speech Processing (ROCLING 2013), (2013) 250-259.

[11] M. A. Ahmed, and M. A. Ahmed, Farmers' Perception and Adaptation to Climate Change: Multinomial Logistic Regression Model Evidence, SSRG International Journal of Economics and Management Studies, 6(10)(2019) 110–119.

Table 4. Calibration Result of the Model

Neologism	Description	Value
자강두천	Fierce competition between two greatly self-esteemed individuals	0.9071
UBD	Measurement of cinema audiences derived from the failed Korean movie	0.3587
뇌절	Continuously repeating a phrase that annoys people	0.6445
아이건좀	The expression indicates denial and concern	0.2645
아이엠그루트	‘I have a lot to say, but stay silent to avoid any conflict	0.6445

Table 5. Prediction Result of the Model

Neologism	Description	Value
자강두천	Fierce competition between two greatly self-esteemed individuals	0.9071
UBD	Measurement of cinema audiences derived from the failed Korean movie	0.3587
뇌절	Continuously repeating a phrase that annoys people	0.6445
아이건좀	The expression indicates denial and concern	0.2645
아이엠그루트	‘I have a lot to say, but stay silent to avoid any conflict	0.6445

.