Topic Detection and Classification in Consumer Web Communication Data

Atsuho Nakayama

Abstract In this study, we examined temporal variation in topics regarding new products by classifying words into clusters based on the co-occurrence of words in Twitter entries. To help identify market trends, analysis of consumer tweet data has received much attention. We collected Twitter entries about new products based on their specific expressions of sentiment or interest. The matrix obtained from the Twitter entries are sparse and of high dimensionality, so we need to perform a dimensionality reduction analysis. We analyzed the matrix using non-negative matrix factorization to reduce the dimensionality. We also clarified temporal variation by using the weight coefficients which show the strength of associations between entries and topics. It is important to consider the temporal variation of these topics when detecting trending topics by classifying words into clusters based on co-occurrence of words.

Atsuho Nakayama
Tokyo Metropolitan University, 1-1 Minami-Ohsawa, Hachioji-shi, 192-0397, Japan
✉ atsuho@tmu.ac.jp

Archives of Data Science, Series A
(Online First)
KIT Scientific Publishing
Vol. 5, No. 1, 2018

DOI: 10.5445/KSP/1000087327/11
ISSN 2363-9881

1 Introduction

The distribution of posts on the internet has been increasing. Today uploading habits have become part of our lifes. For example, sharing photos on social networking sites, instant sharing with smartphones. People use texts and images on the internet to represent their activities, interests and opinions. Despite different demographics, posts of different users contain similar interests. It is important to use consumers' uploading habits on the internet for marketing purposes. Bendle and Wang (2016) demonstrated the application of a topic model in the business field, especially market research, from a conceptual standpoint to solve this problem. They discussed the role and benefits of topic model application to market research. They argued that what is revealed by applying a topic model to user-generated content is a discussion about the competitive structure of the market from the consumer's perspective and the positioning of companies, products and services. In addition, there are studies by Liu et al (2017), and Tirunillai and Tellis (2014) with a focus on the same points. In these researches that focused on product positioning, topic models are applied to user-generated content to extract potential consumer-oriented attributes in an industry, product, or brand. The extracted attributes are mainly used for positioning companies and products. In the marketing field, this is the mainstream research for the application of the topic model. A problem of the analysis of unstructured consumers reviews is the revelation of topics that are represented by the words used to express their experiences.

The purpose of this study is to detect trending topics in online word-of-mouth data with a focus on topics related to new products. This was done by classifying words into clusters based on their co-occurrence. We collected Twitter entries about new products based on their specific expressions of sentiment or interest. Twitter has been spreading in Japan. Twitter users post to represent their activities, interests and opinions. To help identify market trends, the analysis of consumer tweet data has received much attention. In this study, we examined temporal variation in topics regarding new products by classifying words into clusters based on the co-occurrence of words in Twitter entries. Personal concerns are influenced by product strategies, such as marketing communication strategies, and thus change over time. It is important to consider the temporal variation in trending topics when detecting such trending topics by classifying words into clusters based on co-occurrence of words in Twitter entries. We chose keywords

representing various topics from Twitter entries and tracked the weekly variation in these topics. We then classified the words extracted from Twitter data using topic model such as non-negative matrix factorization (NMF) to reduce the dimensionality (Lee and Seung, 2001).

2 The Data

The text data of Twitter entries regarding certain product names were searched and collected at 5-minute intervals. We created a system for data cloning that was programmed in Ruby. Due to changes in specifications of the system, our Twitter API is not currently operational. However, until this change, we were able to easily collect Twitter entries by using the R package "twitterR". Similar libraries are updated and released in many programming languages. We searched for Twitter entries regarding a new brand of inexpensive, beer-like beverage named "金のオフ" (Kin no Off), produced by Sapporo Breweries Ltd., and we collected 4622 tweets from September 2nd, 2011 through May 18th, 2012. This is the same data as in Nakayama (2017). Our reasons for focusing on Twitter entries regarding new beverage products were twofold: It is a useful means of diminishing the effect of past product strategies such as merchandising and advertising, and we have found it particularly easy to evaluate time series variation in personal concerns for newly released beverage products.

In this study, we looked for topics associated with new products by classifying words into clusters based on the entry × word matrix of Twitter entries. To detect topics more easily, we tokenized each tweet message that was written in sentences or sets of words. However, one of the most difficult natural language-processing problems in Japanese is tokenization. This is referred to as the "wakachigaki" problem. In most Western languages, words are delimited by spaces and punctuation. So, the white spaces, numbers and stopwords can be eliminated to keep clean linguistic content if we use the `tm` package of R to analyze English text.

In Japanese, words are not separated by spaces. An example is the following sentence "私は新商品を購入した" (watashihashinshouhinwokounyuushita, translated: "I bought a new item") which shows no spaces or separation symbols between the Japanese words. We used morphological analyses such as tokenization, stemming, and part-of-speech tagging to separate the words. In our study, we used the Japanese morphological analyzer ChaSen to separate

words in passages and to distinguish all nouns, verbs, and adjectives. This procedure excludes function words such as particles and auxiliary verbs that can be regarded as stop words. ChaSen is a fast, customizable Japanese morphological analyzer that takes the form of a hierarchical structure. It is designed for generic use, and can be applied to a variety of language-processing tasks – a detailed discussion of ChaSen can be found in Kudo et al (2004).

Next, we selected keywords representative of our chosen topics. To better understand topic characteristics, it was important to establish criteria to choose appropriate words representing temporal variation. We performed a statistical analysis based on the complementary similarity measure (CSM; Sawaki and Hagita, 1996). CSM has been widely applied in the area of character recognition, and was originally developed for the recognition of degraded machine-printed characters. To construct appropriate word-set topics each week, we estimated the associations within word pairs. CSM is able to measure the inclusion relation between weeks i and j to recognize characters and identify word trends on a weekly basis. Given the following table of data (see Table 1), CSM is defined as follows:

$$\text{CSM}\,(\text{Week } i,\ \text{Week } j) = \frac{(ad - bc)}{\sqrt{(a + d)\,(b + c)}}\,. \tag{1}$$

Table 1: Example to demonstrate the method used to calculate CSM.

	Week i	Week j
Frequency of the word X	a	b
Frequency of words other than X	c	d

Chi-square values have often been used to estimate the relation between two words. The formulas for CSM and chi-square are quite similar. However, the chi-square analysis is more likely to select words occurring with low frequency compared to the CSM method when analyzing data that contain a large spread in the occurrence frequency of words. Certain words occurred only rarely, whereas others occurred quite frequently in the text data of Twitter entries used in this study. Thus, the frequency of occurrence of some words was hundreds of times larger than that of others. For this reason, we decided to use CSM in this study. We collected the words receiving the top 10 CSM scores each week, and retained words with a total selection frequency of eight or more. The CSM score depends

on word frequency. Thus, it was possible for words with low total frequency of occurrence to be selected as distinct words during a particular week, provided that the words occurred frequently that week. We removed all entries that did not include any of these words. The data comprised 4232 entries × 358 words. The data showed co-occurrences among 358 words in the selected entries.

The words related to the launch of new products, the words related to a TV commercial about new products, and the words related to purchasing or tasting new products are extracted as feature words in a few weeks since launch. The words related to a TV commercial about new products will not be extracted as feature words as the GRP (Gross Rating Point) will greatly decrease in about 2 months after the launch of new products. On the other hand, the words “drink”, “delicious”, “tasty”, “rich”, and “bitter” related to tasting the product are extracted. . The GRP is widely used as indicator to measure of the size of an advertising campain by a specific period. Let cumulative arrival rate of ads be R and average freguency of ads be F. GRP is represented the following Equation 2:

$$R \times F = GRP. \tag{2}$$

The GRPs were provided by Video Research Co., Ltd. (Japan); for more details on the GRP, see VR Digest (`https://www.videor.co.jp/digestplus/tv/2017/05/7917.html?wovn=en`). Therefore, it is considered that the words can be extracted to reflect the difference of the words appearing in each week. For example, the words such as “Sapporo”, “Nagasaku”, “cute”, “purchase”, “tasty”, and “new” are extracted as feature words on a launch week of the new product. “Nagasaku” shows the name of the actress who appeared in the “CM” (TV commercial), “cute” means the impression of the impression about the actress by watching the TV commercial of the product, “Sapporo” is the name of manufacturer that released the product, “Buy” and “Taste” represent that consumer purchased or tasted the new product, and "new" indicates the relationship with the launch of a new product. Therefore, words that reflect consumer behavior and impressions are extracted. “Nagasaku” is the most frequent at 307 times and “new” is the least frequent at 17 times in these words. “Nagasaku” appeared 1221 times in all periods and this is 3 % of all words. On the other hand, “new” appeared 36 times and it is only 0.1 % of the total. By using the CSM for word extraction, it is possible to extract words that appear in a specific period but have a low appearance frequency as feature words.

3 The Analysis

3.1 Analysis of Topic Classification

The entry × word matrix obtained from the Twitter entries was sparse and of high dimensionality, so it was necessary to perform a dimensionality reduction analysis. We employed some excellent computing resources to help analyze the highly dimensional and sparse matrices. In addition, these matrices often contained noise, making it difficult to uncover the underlying semantic structure. Because of these difficulties, we found it necessary to implement dimensionality reduction. To reduce dimensionality, procedures such as Latent Semantic Analysis (LSA) or Latent Semantic Indexing (LSI; Deerwester et al, 1990) and Probabilistic Latent Semantic Analysis (PLSA) or Probabilistic Latent Semantic Indexing (PLSI; e.g. Hofmann, 1999) are often applied. LSA reduces the dimensionality of the entry × word matrix by applying a singular value decomposition (SVD), and it then expresses the result in an intuitive and comprehensible form. However, it can take a long time to perform LSA on a large matrix. In PLSA, a probabilistic framework is combined with LSA. This method uses mixture decomposition (the convex combination of aspects), which has a well-defined probability distribution. The factors have clear probabilistic interpretations in terms of the distribution of mixture components. We analyzed the entry × word matrix using non-negative matrix factorization (NMF) to reduce the dimensionality (Lee and Seung, 2001). Similar to principal component analysis (PCA), NMF consists of positive coefficients in linear combination. The computation of NMF is based on a simple iterative algorithm, which is particularly useful for applications involving large, sparse matrices. Ding et al (2006) have shown that both NMF and PLSI (PLSA) optimize the same objective function, ensuring that the use of NMF and PLSI are equivalent. NMF is used for dimensionality reduction as follows:

$$V_n \approx W_n \times H_r, r < \frac{n \cdot m}{(n+m)} \tag{3}$$

The matrix V consists of non-negative data, such as that in an entry × word matrix. The matrix W contains non-negative basis vectors and shows the strength of associations between words and topics. The matrix H contains non-negative coefficients and shows the strength of associations between entries and topics. We can detect topics involving new products using the basis vector coefficients. The results are conceptually similar to those of Principal Component Analysis

(PCA), but the basis vectors are non-negative. Here, the original data are represented purely through additive combinations of the basis vectors. This characteristic of NMF, i.e., data representation based on additive combinations, is effective because it suggests the intuitive notion of combining parts to form a whole. NMF computation is based on this simple iterative algorithm, and it is very efficient for applications involving large matrices.

Note that throughout this section, Japanese words will be followed by their English translations in parentheses. We classified the words extracted from the tweet data regarding a new brand of inexpensive, beer-like beverage named "金のオフ?" (Kin no Off) produced by Sapporo Breweries Ltd. We implemented NMF to reduce dimensionality using the R package "NMF" based on Lee's model (Lee and Seung, 2001). "Kin no Off" contains 50 % less purine and 70 % less carbohydrates than other inexpensive, beer-like beverages. It is thus classified as a third-category beer, containing ingredients such as corn, soybeans, and peas rather than malt for the purpose of price reduction. For Japanese taxation purposes, brewed malt beverages in Japan fall into one of three categories: beer, Happoshu, or third-category beer. Alcoholic beverages made from malt are classified as beer if their malt content exceeds 67 %. If a beverage contains less than 67 % malt content, it falls under the tax category of Happoshu. Japanese breweries have produced even lower-taxed and non-malt brews made from soybeans and other ingredients, which do not fall under either of these classifications. These lower-taxed, non-malt brews, referred to by the mass media as third-category beers, were developed to compete with Happoshu.

Lee's model is an algorithm based on Euclidean distance that uses simple multiplicative updates. We determined that the maximum number of topics was 10, and the minimum as 4. In this analysis, eight topics are discussed for interpretation purposes.

Table 2 lists the eight topics and shows the top 10 heavily weighted words in the basis vector W. Spellings using the Roman alphabet as well as English translations of the Japanese words are also shown in Table 3. From results such as these, we are able to identify the one or two words that are most heavily weighted. Each topic consists of a small, core set of words to convey intention with short sentences. We can detect the prevalence of certain topics based on observations of which words are most heavily weighted.

We were able to divide the eight topics into three groups. One was the review topics, which consisted of Topics 1, 3, 7, and 8. Topic 1 was the review

containing a link to an external website and product images, Topic 3 was the review of purchasing behavior and information about the new product, Topic 7 was the review of the brewery's release of the new product, and Topic 8 was the review of experiences actually drinking the product. The second group was the topics associated with advertising, which consisted of Topics 2, 5, and 6. Topic 2 was about advertisements on the train, Topic 5 was about TV commercials, and Topic 6 was concerned with performers in TV commercials. The third group consisted only of Topic 4. Topic 4 was not associated with inexpensive beer-like beverages, and the product name used as a keyword to extract Twitter entries that occurred in a different context.

Table 2: The eight topic results and the top 10 weighted Japanese words in the basis vector W (reprints of Table 3 of Nakayama (2017, p. 168); 1/3).

Japanese	Roman Alphabet	English Translation	Weight
		Topic 1	
http	http	http	0.48
オフ	ofu	off	0.02
更新	koushin	update	0.02
ブログ	blog	blog	0.02
良い	yoi	good	0.02
発売	hatubai	release	0.01
なう	nau	now	0.01
ひる	hiru	daytime	0.01
新CM	shinCM	new TV commercial	0.01
新発売	shinhatubai	new release	0.01
		Topic 2	
可愛い	kawaii	cute	0.28
広告	koukoku	advertisement	0.21
見る	miru	see	0.08
良い	yoi	good	0.05
電車	densha	train	0.05
永作	Nagasaku	Nagasaku	0.04
電車内	denshanai	on the train	0.03
人	hito	people	0.03
ー	ー	ー	0.02
ω	ω	ω	0.02
		Topic 3	
ビール	biiru	beer	0.13
買う	kau	purchase	0.08
オフ	ofu	off	0.07

Table 2: The eight topic results and the top 10 weighted Japanese words in the basis vector W (reprints of Table 3 of Nakayama (2017, p. 168); 2/3).

Japanese	Roman Alphabet	English Translation	Weight
	Topic 3		
美味しい	oisii	delicious	0.06
プリン体	purintai	purine	0.04
味	azi	taste	0.04
50％	50％	50％	0.04
発泡酒	happoushu	low-malt beer	0.04
上手い	umai	tasty	0.03
糖質70	toushitu70	carbohydrate 70	0.02
	Topic 4		
ＲＴ	RT	RT	0.21
なう	nau	now	0.04
w	w	w	0.03
金	kin	Friday	0.02
オフ会	ofukai	alcoholic party	0.02
予定	yotei	schedule	0.02
いる	iru	stay	0.02
下さる	kudasaru	do	0.02
お願い	onegai	please	0.02
宜しい	yoroshii	kind regards	0.02
	Topic 5		
CM	CM	TV commercial	0.51
見る	miru	see	0.06
ー	ー	ー	0.03
可愛い	kawaii	cute	0.03
出る	deru	perform	0.03
似る	niru	resemble	0.03
やる	yaru	do	0.02
好き	suki	like	0.02
パフ	Pafu	Puff	0.01
曲	kyoku	music	0.01
	Topic 6		
永作	Nagasaku	Nagasaku	0.56
可愛いすぎる	kawaiisugiru	way too cute	0.12
見える	mieru	appear	0.03
ポスター	Posuta	poster	0.02
大島優子	Oshima Yuuko	Yuuko Oshima	0.02
似る	niru	resemble	0.02
好き	suki	like	0.01
車内広告	shanaikoukoku	advertisement on the train	0.01
男装	dansou	dressing as a man	0.01
ひる	hiru	daytime	0.01

Table 2: The eight topic results and the top 10 weighted Japanese words in the basis vector W (reprints of Table 3 of Nakayama (2017, p. 168); 3/3).

Japanese	Roman Alphabet	English Translation	Weight
	Topic 7		
サッポロ	Sapporo	Sapporo	0.48
上手い	umai	tasty	0.02
味	azi	taste	0.01
発泡酒	happoushu	low-malt beer	0.01
出る	deru	release	0.01
金麦	Kinmugi	Kinmugi	0.01
発売	hatubai	release	0.01
ひる	hiru	daytime	0.01
良い	yoi	good	0.01
こだわる	kodawaru	pursue	0.01
	Topic 8		
飲む	nomu	drink	0.43
寝る	neru	sleep	0.02
見る	miru	see	0.02
味	azi	taste	0.02
ー	ー	ー	0.02
美味しい	oishii	delicious	0.02
好き	suki	like	0.01
笑	wara	laugh	0.01
なう	nau	now	0.01
いる	iru	stay	0.01

Topics 1, 3, 7, and 8 are all based on reviews, though in various ways. In Topic 1, the words associated with the review containing a link to an external website and product images are heavily weighted. The most heavily weighted word was "http", so it is the core word of Topic 1. Other words with comparatively large weights, ranking within the top 10, were often found along with the core word in tweets. Some Twitter entries were posted containing links to the external website and in-line product images, as well as phrases such as

– "the new product 'Kin no Off' was released, and I updated my blog about it",
– "the new TV commercial for the new product 'Kin no Off'was broadcast" or
– "'Kin no Off' is a new release and it tastes good".

Some users posted links to external websites, such as their own blogs or the manufacturer's homepage. Others added in-line product images to their tweets.

We, therefore, believe that it would be possible to infer the topic of these tweets, namely reviews containing a link to an external website as well as product images, solely from the most heavily weighted words of Topic 1.

In Topic 3, the words associated with the review of purchasing behavior and information about the new product were heavily weighted. The most heavily weighted word was "ビール" (beer), followed by "買う" (purchase / buy). These words are the core words of Topic 3. Other words with comparatively large weights, ranking within the top 10, were often found along with the core words in tweets. "Kin no Off" contains 50 % less purine and 70 % less carbohydrates than other third-category beers, and is thought to be a healthier product. We believe that these features of the new product can be inferred from the list of heavily weighted words of Topic 3. Actual Twitter entries corresponding to this topic include

- "I bought the third-category beer named 'Kin no Off,' and it features 50 % reduced purine and 70 % reduced carbohydrate",
- "the catch-phrase of the third-category beer named 'Kin no Off' is that it is delicious, though the purine and carbohydrate are reduced, so we should purchase it if its taste is as delicious as that of low-malt beer or especially normal beer" and
- "the features of the third-category beer named 'Kin no Off' include are 50 % reduced purine and 70 % reduced carbohydrate, and it is as tasty as normal beer".

To repeat, Topic 3 is reflected in tweets concerning purchasing behavior and information about the new product.

In Topic 7, the words associated with the review of the release of the new product from Sapporo BreweriesLtd. are heavily weighted. The most heavily weighted word is the Brewery's name, "サッポロ" (Sapporo), and it is the core word of Topic 7. Other words with comparatively large weights, ranking within the top 10, were often found along with the core word in tweets. "金麦" (Kinmugi) is a rival third-category beer. Further Twitter entries include

- "Sapporo 'Kin no Off' is tasty, and the taste is better than other low-malt beers, so I think the materials to make it were selected carefully" and
- "I made a trial purchase of Sapporo 'Kin no Off' that had been newly released, and its taste is good".

To repeat, Topic 7 is associated with the release of the new product from Sapporo BreweriesLtd. and reviews of its taste.

In Topic 8, the words associated with reviews of drinking the product have the heaviest weight. The most heavily weighted word is "飲む" (drink), and it is the core word of Topic 8. Other words with comparatively large weights, ranking within the top 10, were often found along with the core word in tweets. Some examples of Twitter entries associated with Topic 8 include

- "I like to drink 'Kin no Off'", "personally, it is my very favorite taste",
- "I will sleep well after drinking 'Kin no Off' because I am tired today" and
- "I drank 'Kin no Off,' and it was more delicious than other third-category beers".

To repeat, Topic 8 is associated with reviews of product consumption. Topics 2, 5, and 6 are associated with advertising. Topic 2 regards advertisements on the train, Topic 5 is associated with a TV commercials, and Topic 6 concerns a TV commercial performer. In Topic 2, the words associated with advertisements on the train have the heaviest weight. The most heavily weighted word is "可愛い" (cute), followed by "広告" (advertisement). These words are the core words of Topic 2. Other words with comparatively large weights, ranking within the top 10, were often found along with the core words in tweets. Hiromi Nagasaku (永作 博美), a popular Japanese actress, appeared in the advertisements on the train. We believe that it would be possible to infer this by observing the top words of Topic 2. In Topic 5, the words associated with advertising have the heaviest weight. The most heavily weighted word is "CM" (TV commercial), and it is the core word of Topic 5. Other words with comparatively large weights, ranking within the top 10, were often found along with the core word in tweets. The Twitter entries generally contained positive feedback with regards to the performer in the TV commercial. The song "Puff, the Magic Dragon" played during the TV commercial, and Twitter entries addressing the music were also posted. We believe that the top words of Topic 5 reflect tweeters' impressions of the performer and music in the commercial. In Topic 6, the words associated with the performer in the advertisement are most heavily weighted. The most heavily weighted word is "永作" (Nagasaku), the name of the performer, followed by "可愛すぎる" (way too cute). These words are the core words of Topic 6. Other words with comparatively large weights, ranking within the top 10, were often found along with the core words in tweets. In the advertisement, Nagasaku is dressed as a man. Twitter entries regarding this topic have generally been positive, and have included phrases such as

– "I like the TV commercial performer 'Nagasaku'",
– "the TV commercial performer 'Nagasaku' dressed as a man in the advertisement is cute" or
– "the TV commercial performer 'Nagasaku' resembles 'Yuuko Oshima'" (a popular Japanese actress and singer).

Therefore, we believe that tweeters' general impressions of Nagasaku in the commercial can be inferred by observing the list of top words for Topic 6.

3.2 Analysis of Topic Transition Patterns

Personal concerns are influenced by new product strategies, such as marketing communication strategies, and they change over time. It is important to consider the temporal variation of these topics when detecting trending topics by classifying words into clusters based on the co-occurrence of words.

On September 22nd 2011, the new brand of inexpensive, beer-like beverage was launched, and TV commercials for the product began running that week. A peak in Twitter entries was reached a few weeks after the product release. The number of entries per week slowly decreased after this peak. The gross rating point declined during the 2 months following the release, and the number of entries also decreased. Small peaks, however, were triggered by the release of new TV commercials. These data show how the weekly number of Twitter entries exhibited temporal change. Thus, to understand topic characteristics, it is important to consider temporal variation in trending topics and to establish criteria to select appropriate words that are representative of such temporal variation.

We clarified temporal variation by using the weight coefficients H which show the strength of associations between entries and topics. The data comprised 4232 entries $\times$ 8 topics. In order to capture the tendency of entries, we analyzed the weight coefficients H by PCA and t-SNE (t-Distributed Stochastic Neighbor Embedding; van der Maaten and Hinton, 2008). t-SNE is a technique for dimensionality reduction that is particularly well suited for the visualization of high-dimensional datasets. Contrary to PCA, t-SNE is not a mathematical technique but a probabilistic one. t-SNE minimizes the divergence between two distributions: A distribution that measures pairwise similarities of the input objects and a distribution that measures pairwise similarities of the

corresponding low-dimensional points in the embedding. Essentially, what this means, is that it looks at the original data that is entered into the algorithm and looks at how to best represent this data using less dimensions by matching both distributions. Hinton and Roweis (2003) proposed a way of converting a high-dimensional dataset into a matrix of pairwise similarities and presented a technique for visualizing the resulting similarity data. t-SNE is efficiently able to capture the local structure of the high-dimensional data and show global structure such as the presence of clusters at several scales. The technique is based on SNE (Stochastic Neighbor Embedding; Hinton and Roweis, 2003), that is much easier to optimize, and produces significantly better visualizations by reducing the tendency to crowd points together in the center of the map.

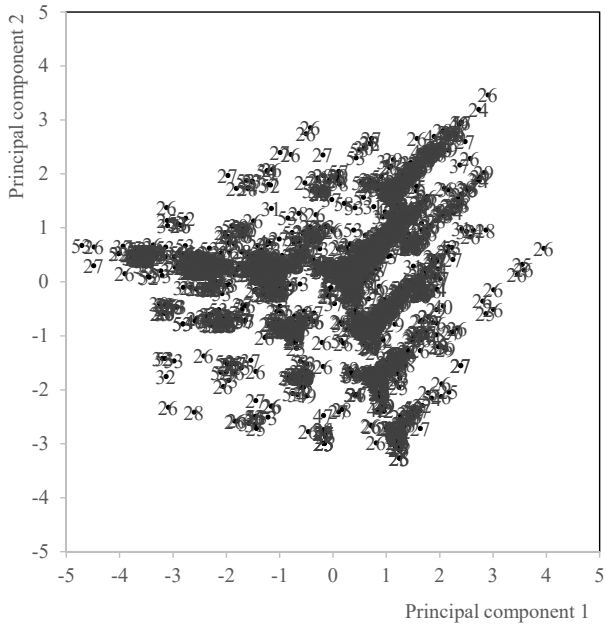

Figure 1: Two-dimensional configuration of the results obtained from analysis of PCA.

To reveal the structure of entries, we analyzed the weight coefficients H by PCA and t-SNE. Figure 1 shows the configuration obtained from analysis of PCA and Figure 2 shows the configuration obtained from analysis of t-SNE. In these figures, the number represents weeks, numbered from 24 to 60, where 24 was

the first week of data collection, and 60 is the last. We collected tweets from September 2nd, 2011 through May 18th, 2012. Each point corresponds to an entry and the label shows the week when each entry was posted. Figure 1 was not able to clearly show the difference between the features of entries. We could not capture the tendency of entries from the figure. The cumulative contribution rate is about 35 % to the second principal component. PCA could not reveal the relationship between entries.

In Figure 2, the shades of label colors indicated differences of weeks. The lighter shade of the labels were the week immediately after the release of the new product. The darker the shade of the label, the more time had elapsed since the release of the new product. Figure 2 reveals that several groups of entry exist. Some groups consist of similar shades. This shows that there were topics posted in a specific period. The others consist of various shades. They show that there were topics posted in multiple periods. These results were thought to be caused by the influence of personal concerns on tweet contents of Twitter users.

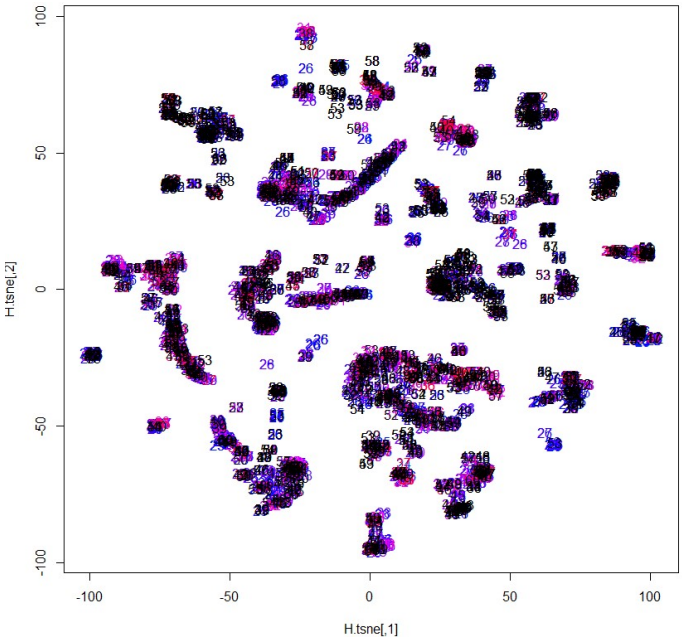

Figure 2: Two-dimensional configuration of the results obtained from analysis of t-SNE.

4 Conclusion

We detected trending topics related to a new product by classifying words into clusters based on the co-occurrence of words in Twitter entries. We were able to divide the eight topics into three groups: Those associated with reviews, those associated with advertising, and those not associated with inexpensive beer-like beverages. Each topic also consisted of a small set of core words to convey intention with short sentences. These topics were further classified by the characteristics of their core words. We detected trends in topics related to new products by using the weight coefficients H which show the strength of associations between entries and topics. We found that there are topics posted in a specific or multiple periods. These results are thought to be caused by the influence of personal concerns on tweet contents of Twitter users. From the results obtained by t-SNE, it was found that time series similarity and dissimilarity of entries existed. However, the reason why such similarity and dissimilarity exist is not sufficiently discussed. We need to discuss this in future study. In the analysis, expressions like "http" and "beer" got high weights. We did not set most words as stop words, because we aimed to capture typical posting trends on Twitter. But it might be better to consider these words as stop words, because this text collection is about beer products. Therefore, we think it is an important challenge to consider the setting of stop words in a future study.

Acknowledgements We express our gratitude to the anonymous referees for their valuable reviews. This work was supported by a Grant-in-Aid for Scientific Research (C) (No. 16K00052) from the Japan Society for the Promotion of Science. We are grateful for financial support from the 45th Yoshida Hideo Memorial Foundation. We wish to thank Video Research Ltd. for allowing us to make use of the GRP data. We are also greatly indebted to Hiroyuki Tsurumi of Yokohama National University and Jyunya Masuda of INTAGE Inc. for their great support and advice in analyzing data.

References

Bendle NT, Wang XS (2016) Uncovering the Message from the Mess of Big Data. Business Horizons 59(1):115–124. DOI: 10.1016/j.bushor.2015.10.001.

Deerwester S, Dumais S, Furnas GW, Landauer TK, Harshman R (1990) Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science 41(6):391–407. DOI: 10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9.

Ding C, Li T, Peng W (2006) Nonnegative Matrix Factorization and Probabilistic Latent Semantic Indexing: Equivalence, Chi-Square Statistic, and a Hybrid Method. In: Proceedings of the 21st National Conference on Artificial Intelligence (AAAI) and the 18th Innovative Applications of Artificial Intelligence Conference (IAAI), Association for the Advancement of Artificial Intelligence (AAAI), The AAAI Press, Menlo Park (USA), pp. 342–347. URL: https://aaai.org/Library/AAAI/2006/aaai06-055.php.

Hinton GE, Roweis ST (2003) Stochastic Neighbor Embedding. In: Advances in Neural Information Processing Systems, Proceedings of the 2002 Neural Information Processing Systems Conference (NIPS 2002), Becker S, Thrun S, Obermayer K (eds), MIT Press, Neural Information Processing Systems Foundation (NIPS), La Jolla (USA), Vol. 15, pp. 833–840. ISBN: 978-0-262025-50-8, URL: https://papers.nips.cc/paper/2276-stochastic-neighbor-embedding.

Hofmann T (1999) Probabilistic Latent Semantic Analysis. In: Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence, Laskey K, Prade H (eds), Morgan Kaufmann, San Francisco (USA), pp. 289–296. URL: https://dslpitt.org/uai/displayArticleDetails.jsp?mmnu=1&smnu=2&article_id=179&proceeding_id=15.

Kudo T, Yamamoto K, Matsumoto Y (2004) Applying Conditional Random Fields to Japanese Morphological Analysis. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Barcelona (Spain), pp. 230–237. URL: https://www.aclweb.org/anthology/W04-3230.

Lee DD, Seung HS (2001) Algorithms for Non-Negative Matrix Factorization. In: Advances in Neural Information Processing Systems, Proceedings of the 2000 Neural Information Processing Systems Conference (NIPS 2000), Leen TK, Dietterich TG, Tresp V (eds), MIT Press, Neural Information Processing Systems Foundation (NIPS), La Jolla (USA), Vol. 13, pp. 556–562. ISBN: 978-0-262122-41-2, URL: https://papers.nips.cc/paper/1861-algorithms-for-non-negative-matrix-factorization.

Liu X, Burns AC, Y. H (2017) An Investigation of Brand-related User-generated Content on Twitter. Journal of Advertising 46(2):236–247. DOI: 10.1080/00913367.2017.1297273.

van der Maaten L, Hinton G (2008) Visualizing Data Using t-SNE. Journal of Machine Learning Research (JMLR) 9:2579–2605. URL: http://www.jmlr.org/papers/v9/vandermaaten08a.html.

Nakayama A (2017) The Classification and Visualization of Twitter Trending Topics Considering Time Series Variation. In: Data Science – Innovative Developments in Data Analysis and Clustering, Palumbo F, Montanari A, Vichi M (eds), Springer Nature, Cham (Switzerland), Studies in Classification, Data Analysis, and Knowledge Organization, pp. 161–173. DOI: 10.1007/978-3-319-55723-6_13.

Sawaki M, Hagita N (1996) Recognition of Degraded Machine-Printed Characters Using a Complementary Similarity Measure and Error-Correction Learning. In: IS&T/SPIE's Symposium on Electronic Imaging: Science and Technology, Document Recognition II, SPIE Proceedings, Society of Photographic Instrumentation Engineers (SPIE), San Jose, Bellingham (USA), Vol. 2422, pp. 491–497. DOI: 10.1117/12.205826.

Tirunillai S, Tellis GJ (2014) Mining Marketing Meaning from Online Chatter: Strategic Brand Analysis of Big Data Using Latent Dirichlet Allocation. Journal of Marketing Research (JMR) 51(4):463–479. DOI: 10.1509/jmr.12.0106.