Analytics of Response for Corporate Twitter Accounts Using Text Mining: Case of Japanese Automotive Manufacturers

Yuta Kitano
Management Science and Social Informatics Program, Department of Informatics
The University of Electro-Communications
1-5-1, Chofugaoka, Chofu-city, Tokyo, Japan
k1930040@edu.cc.uec.ac.jp

Tetsuo Yamada
Management Science and Social Informatics Program, Department of Informatics
The University of Electro-Communications
1-5-1, Chofugaoka, Chofu-city, Tokyo, Japan
tyamada@uec.ac.jp

Kim Hua Tan
Operations Management and Information Systems, Business School
University of Nottingham
NG7 2RD, Nottingham, United Kingdom
Kim.Tan@nottingham.ac.uk

Abstract

Recent years, with the spread of smartphones, Twitter, one of the Social Networking Service (SNS) has become one of the sources of information for people. These new sources of information are also influencing companies' advertising strategies, with companies also having their own Twitter accounts and using them to promote themselves. However, the companies are still searching for Twitter strategies, such as what kind of content will bring a better impression on Twitter users. Therefore, the current status of corporate accounts needs to be investigated and quantitatively analyzed. The purpose of this study is to extract Retweet and LIKE, and then, analyze the strategy of tweeting that brings a better impression on users. Firstly, the target company's Tweet data is extracted, and 5 Japanese automotive companies are targeted as examples of companies. Next, Twitter account data and information obtained per Tweet is analyzed by basic statistics, linear regression, and Text Mining. Finally, from the information obtained in the previous analysis, the elements that are highly correlated with RT, are extracted, and the Tweet strategy is considered.

Keywords
Social Media, Data Science, Data Mining, Automobile Industry, Twitter Strategy

1. Introduction

In today's information society, people can see a lot of information on a daily life. Additionally, with the spread of smartphones (Ministry of Internal Affairs and Communications, 2020), people have got more and more information from their screens. These social changes have also affected corporate advertising efforts. A few years ago, TV commercial was one of the most famous advertising ways. However, corporate TV commerce cost are decreased year by year due to advertising costs in Japan are increased (Dentsu, 2018).
     On the other hand, Social Networking Service (SNS) (Han et al. 2015) is one of the ways to get multiple information for the people due to spread of smartphones. Basically, the main purpose of SNS was promoting

connections between people and people. As time went on, it has become a source of new information for people (Omuka, 2015). Companies have also been influenced by the spread of SNS and the current trend away from TV, and have developed advertising strategies for smartphones as well. One of the strategies is the opening of their own SNS accounts. In other words, companies send various information to users or consumers through SNS. This is now used as an inexpensive advertising tool for Japanese companies. Most of Japanese companies have their own SNS accounts, and there is various usage by companies.

Additionally, Twitter (Ma et al. 2019) which is one of the SNS has a characteristic which spread of tweets across communities has the effect of increasing the final spread of tweets. From such a characteristic, Twitter is suitable as free advertising. People who see the tweet, called followers, can do LIKE or Retweet (RT) the tweet, which means spreading the tweet to their own followers (Twitter Help Center, 2019). Companies use SNS to communicate their activities and their products to users or potential customers. Therefore, a survey of the SNS usage by companies will clarify methods of advertising strategies in the future information age.

Kitano et al. (2019) focused on the content of tweets and used a technique called Text Mining to identify commonalities in the content of tweets. However, they only did word frequency analysis and network analysis. No specific managerial Twitter strategy has been proposed. Ma et al. (2019) used the subject of corporate scandal case to investigate the impact of Twitter. They looked at the accounts of the companies that caused the scandal and the tweets of a number of public users to determine the impact. However, they insist on the use of Twitter from a risk management perspective, focusing on the negative aspects of companies and not on the promotion of corporate activities. Although there are several studies on corporate Twitter accounts, they have not proposed on managerial strategies focused on the number of RT.

This study focuses on RT as an indicator to adapt the use of Twitter to business, and analyzes its relationship with other factors such as LIKE or the number of followers. This study simultaneously considers numerical data from Twitter data and textual data from Tweet contents. In order to achieve the above research objectives, the following research questions (RQ i), ii), iii) are set up.

i)   How many words should be suitable for RT?
        Is there a relationship between the number of characters and the number of RT?

ii)   When does a tweet work for RT?
        Is there a relationship between the timing of a posting tweet and the number of RT?

iii)   What kind of sentences are accepted by users?
        Is there a relationship between the words used in many of the current tweets and the number of RT?

2. Methodology
2.1 Research procedure

This section presents the research procedure from data collection to suggest the Twitter strategies for corporate account. Figure 1 shows the analysis procedure of this study. Before this analysis process, the data must be obtained and processed.

In STEP1, the statuses of the selected accounts are summarized. Furthermore, in order to examine the variance and average of the data, the basic statistic of the objective function, the number of RT, are obtained.

Next, STEP2 focuses on Twitter's feature of having a character limit, so that a relationship between the number of characters and the number of RT are examined. Additionally, it also reveals the relationship between LIKE and RT, which are impressions on tweets.

In STEP3, the timing of tweet is divided into three categories, "Month", "Day of the week" and "Hour". Each timing is examined in relation to RT.

Finally, the content of the tweets should also be noted, as only the number of characters cannot be used to examine the impact of a tweet. Thus, Text Mining reveals what words are used in the content of a tweet. Text Mining reveals what words are used in the content of a tweet.
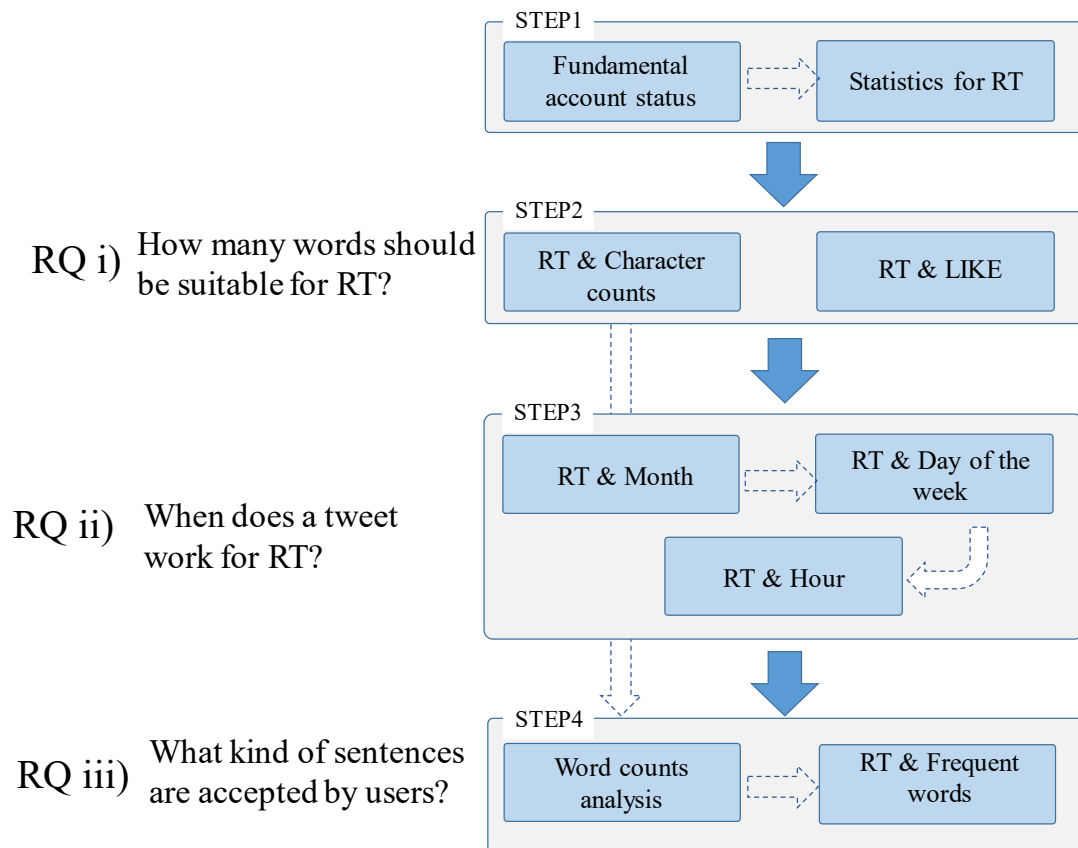
Figure 1. Analysis procedure

2.2 Measurement and statistical analysis for Twitter

It is necessary to use an analytical method suitable for the purpose of Twitter analysis in order to find the RQs. In RQ i), to clarify the relationship between RT and character count and RT and LIKE, the correlation used as an indicator. In RQ ii), the number of tweets and the average number of RT at each of the three timing variables, "Month", "Day of the week", and "Hour", can be found. In RQ iii), at first, the content of the tweets is divided by Text Mining to determine the number of appearances of the words, and next, using the calculated frequency of occurrences as an explanatory variable, the relationship between the tweets and RT is determined by multiple regression analysis.

2.3 Text Mining (TM)

Text Mining (TM) is type of data mining for character strings (Tse et al. 2016). Sentence data is divided into words and phrases and the frequency, word correlation, and time series are analyzed. Since TM is an analytical method for natural languages, it depends on the type of language used. In this study, the tweet data written in Japanese language, and thus, it is necessary to know the difference between Japanese and English. Originally, English has a space between words, in contrast, Japanese has no space between words and it is necessary to separate them by meaning. After dividing the words into meaningful words by using TM software (NTT DATA Mathematical Systems, 2021) a word counts analysis is generally performed.

Word counts analysis (Cohen et al. 2005) examines which words appear frequently in sentences or text to find hot topics and trends. The total number of words and tweets varies depending on the data covered. Therefore, a value called appearance rate in Eq. (1) is used to compare the companies by normalizing the data (Nakamura et al. 2015). Initially, the word frequency for each company is calculated and then the ranking of all 5 companies are created.

$$Appearance\ rate(\%) = \frac{Number\ of\ appearance}{Number\ of\ extracted\ tweets} \times 100 \quad (1)$$

2.4 Used case

In this study, 5 Japanese automotive companies are selected as examples of companies that have a Twitter account due to possessing enough followers and products to advertise with. Additionally, the Japanese automotive industry has one character which is higher advertising costs that others industry (TOYOKEIZAI ONLINE,2017). Thus, they are expected to need a more effective advertising method.

Table 1 shows the corporate Twitter accounts' status used in this study. The most recent year of data, 2019, is used as the target year. At first, the difference between the largest and smallest number of tweets over the year is about 1,000 in this data set. These tweets data include tweets in response to an account called Replies. Next, the number of followers who see the tweets on the corporate account are all exceeded 100,000. In other words, this is presumably similar to the extent to which information can be transmitted.

Table 1. The corporate accounts' status and RT averages

Companies	Year	Followers (04.Sep.2020)	Number of Tweets
Mazda	2019	118532	332
Honda	2019	235609	1383
Toyota	2019	267030	965
Nissan	2019	277714	792
Subaru	2019	166667	335

3. Numerical analysis of corporate accounts and their tweets (RQ i))
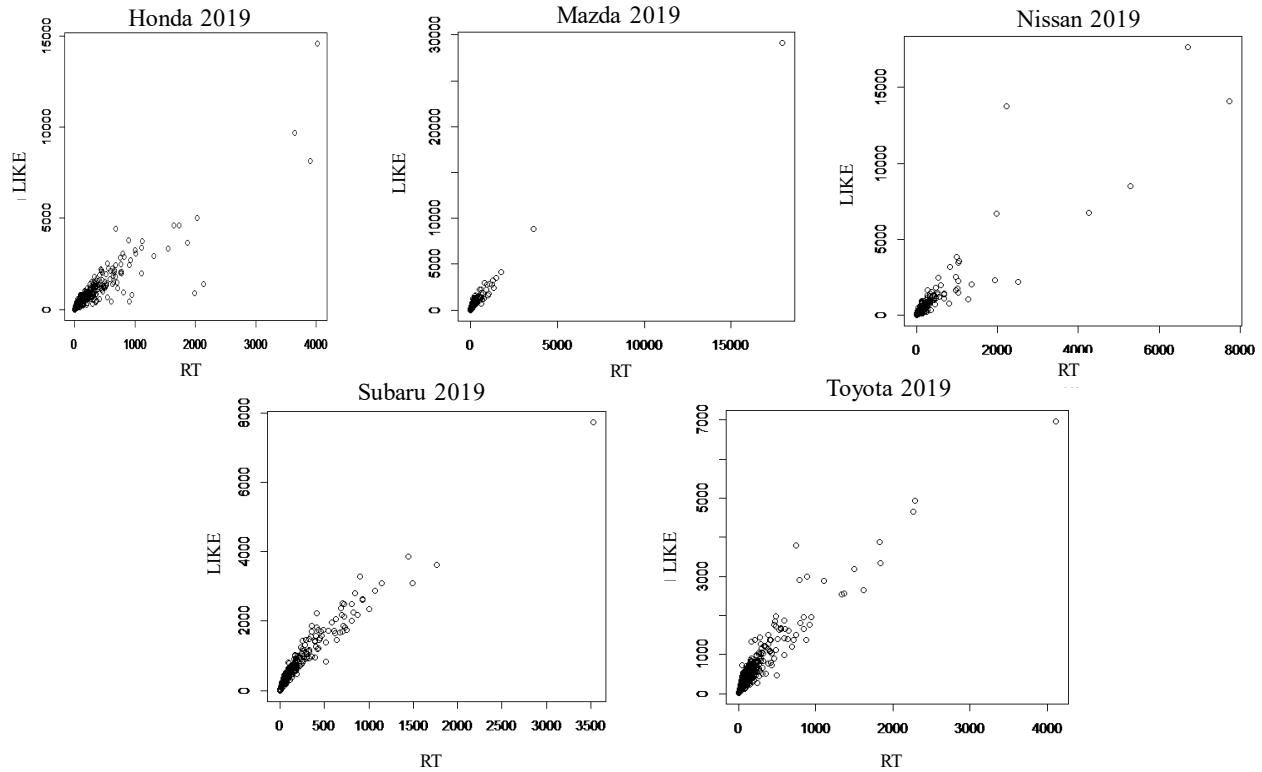### 3.1 The basic status for RT

Table 2 shows the trends in RT from the companies that have received the most attention from this study. There is a large difference in the maximum number of RT relative to the number of followers. Specially, Subaru has more followers than Mazda, however, Mazda has more maximum RT. It can be said that the maximum number of RT on a company account does not necessarily depend on the number of followers. Additionally, Mazda's SD is larger than any other 4 companies' ones due to the maximum retweet of about 17,000. This means that the tweets have been spread more than the usual tweets.

To visualize the normal and sent tweet groups, as in this example, a scatter plot was created. Figure 2 shows a scatter plot of RT and LIKE for each company. As shown in Figure 2, some tweets were found to be unusually retweeted compared to other normal groups of tweets. For example, the top of RT in Mazda is extremely more than the average group.

As a result, the maximum number of retweets or SDs show how much information was spread compared to the average tweet in the automotive industry.

Table 2. The basic status for RT

Companies	Number of Tweets	Max of RT	Mean of RT	Median of RT	SD of RT
Mazda	332	17954	253.7	128.0	1016.9
Honda	1383	4019	105.9	37.0	261.2
Toyota	965	4116	104.7	48.0	238.8
Nissan	792	7743	127.2	40.0	477.4
Subaru	335	3532	183.1	76.0	313.7

3.2 Analysis of the relationship between RT and LIKE and Character counts

To know the relationship between RT and other factors, this study also compares other information contained in the tweets to RT. Information such as date and time of posting, number of RT, number of LIKE, and number of characters in a single tweet.

Table 3 shows correlation between RT and LIKE and correlation between RT and character counts. As a result, a strong positive correlation of more than 0.9 is identified for RT and LIKE. This is likely because both RT and LIKE have positive connotations for the tweet. In contrast, the correlation between RT and character count show little or no correlation close to 0. From these results, the only primary data are included in the tweets that was directly related to RT was LIKE. RT and LIKE have a strong positive correlation in automotive corporate accounts. However, LIKE is a variable that continuously increases during a certain amount of time, just like RT, thus, it is not able to control LIKE when are tweeted. On the other hand, it turns out that the number of characters is not directly related to RT. This tweet data also included URLs and links; therefore, it is not a pure character count. To improve them, it should convert dates to categorical data as well as the number of characters taken out of URLs and the types of words in the tweets. However, there is still the possibility that Twitter, with its character limit, has a connection to RT on whether it has more or less characters.

Next, we have another RQ ii) which is "When does a tweet work for RT?". Thus, the next section should take the information obtained at the time of Tweet generation as a variable when the actual Tweet strategy are considered.

Figure 2. Scatterplots of RT and LIKE for each company

Table 3. The results of correlation between RT and other variables

Companies	Correlation RT*LIKE	Correlation RT*Character counts
Mazda	0.986	-0.020
Honda	0.919	-0.037
Toyota	0.939	-0.049
Nissan	0.913	0.015
Subaru	0.961	0.178

4. Analysis of the relationship between RT and time series data (RQ ii))

In this section, in order to know when a tweet is sent out, the relationship between data on the time a tweet was posted and RT are examined. Data from three time periods, "Month", "Day of the week", and "Hour", that may be related to RT are targeted.

As a result of Figure 3, 4 of the 5 companies had the highest number of tweets in October. This is likely due to the fact that "Tokyo Motor Show 2019", an important event in the Japanese car industry, held in October. Moreover, in Subaru's example, the number of tweets in January is higher due to the number of tweets about their products at the January event. From these findings, the number of tweets during event-related times is higher than in other months. On the other hand, the average change in RT and the number of tweets per month is focused on. Seen Honda's March one and Mazda' April one, there is a maximum of RT average. The reason for this would be that the most retweeted tweets of the year are included in that month. In other words, months of many tweets and average RT are not necessarily related.

As a result of Figure 4, basically, the number of tweets on weekdays is higher than the number of tweets on holidays. Additionally, all five companies were found to have the highest number of tweets on Friday. The content of the tweets was not specific to Fridays, however, across the company, Fridays were found to be the most tweeted about. On the other hand, the average number of RTs was characterized differently by different companies. For example, the average RTs of Toyota and Honda did not change much regardless of the day of the week, however, the average RTs of the other three companies varied widely by day of the week. Thus, some companies may reduce the productivity of RTs on certain days of the week were found.

As a result of Figure 5, all five companies have a high concentration of tweets between about 8am and 6pm, which is basically corporate business hours. Totally, all five companies were found to have a relatively high number of tweets between 5pm-6pm. However, in some cases, such as Mazda and Subaru, they are concentrated at certain times of the day. On the other hand, the average RTs did not have a common peak time, however, the examples of Nissan and Mazda had higher numbers for both tweets and average retweets at around 11am. Like Mazda and Nissan, the time when many tweets are tweeted and the average number of RT is better timing when a lot of information spread.

In order to find out more of the relationship between time and RT, the correlation between average RT and each time parameter. Additionally, the same verification was done with the top 30 RTs of each company to check not only the total tweets but also the top trends. Table 4 shows the results of correlation between average RT and the parameters of time. At first, a strong correlation did not appear in the results for all tweets. However, the correlation was more pronounced in the Top 30 results than in all the tweets, although the data is more biased. Especially, in Nissan, the correlation is high in "Weekday" and "Hour". This is expected to be appropriate for the timing of the tweets and the average of RT.

From these results, trends in the number of tweets were identified for each data set. In hours, the coincidence of the time when most tweets are tweeted and the average RT are comparatively high is productive tweeting. This is one factor that indicates that not only the content of the tweet but also the timing of the post is important. Additionally, strong positive correlations were also partially identified in the Top30 RT. This at least indicates that post timing is also important in getting a lot of RT.
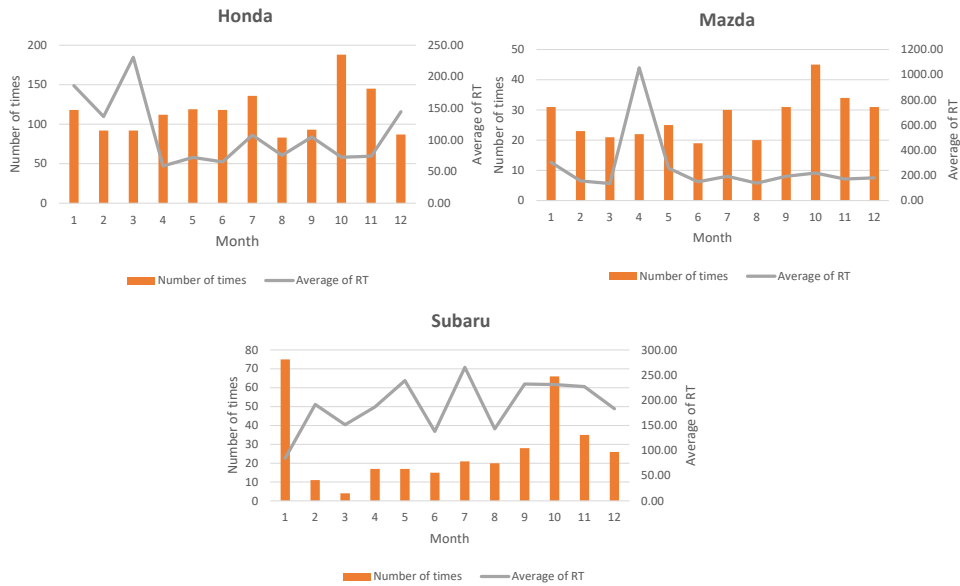
Figure 3. The relationship between average RT and number of tweets per month
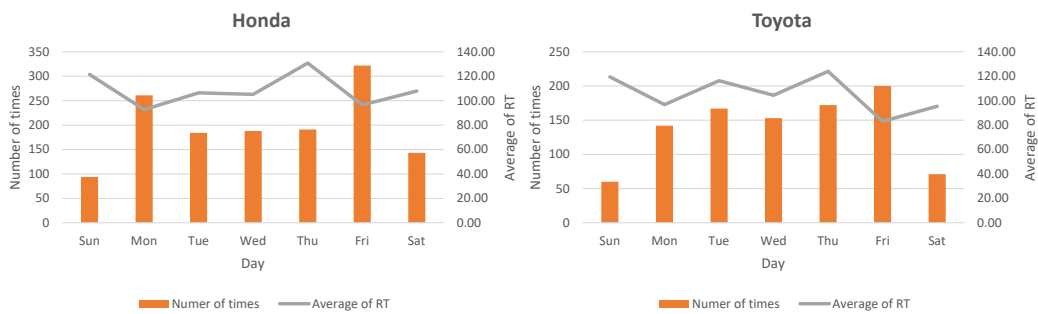

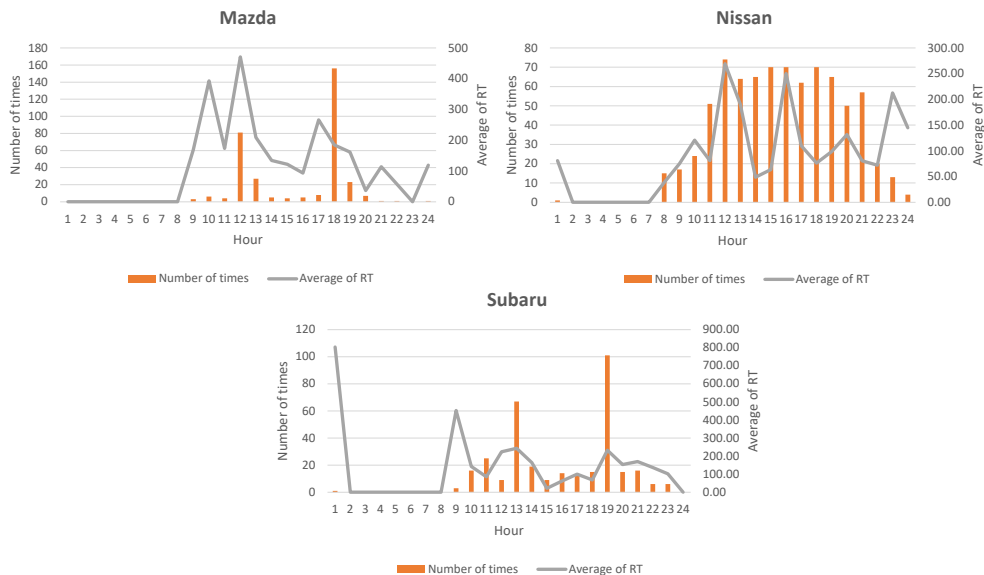Figure 4. The relationship between average RT and number of tweets per day of the week

Figure 5. The relationship between average RT and number of tweets per hour

Table 4. The results of correlation between average RT and the parameters of time

Number of Tweets	Compnies	Month	Weekday	Hour
All	Mazda	-0.146	-0.513	0.456
	Honda	-0.383	-0.648	-0.013
	Toyota	-0.317	-0.190	-0.095
	Nissan	-0.315	-0.282	0.552
	Subaru	-0.175	0.622	0.177
Top30RT	Mazda	0.189	0.247	0.680
	Honda	0.632	-0.857	0.534
	Toyota	0.661	-0.095	0.652
	Nissan	0.238	0.862	0.876
	Subaru	0.381	0.701	0.744

5. Analysis of tweet content (RQ iii))
5.1 Word counts analysis

In the previous chapter, the numerical data of tweets were focused on. Therefore, in this section, TM were used in order to reveal the words tweeted in tweets.

Table 5 shows the results of the word counts analysis. This table is arranged in order of word frequency. From the appearance rate, the content of the tweet can be guessed at. Some of the words indicated the content of the tweets. For instance, the words "Holding" or "Introduction" often mean product and event announcements. Moreover, a word Subaru's "Holding" rate was about 20%. This means that 20% of their tweets were used to announce their events. In addition, the word "Please look" almost included links to websites and videos, which means directing people to informational media other than Twitter or getting information to watch video.

As a result, based on the frequency of the words, it is possible that the words included could predict the content of the tweet. Additionally, how this word specification trend affects RT needs to be examined. The next step is to analyze the relationship between these frequent words and RT.

Table 5. The results of the word counts analysis

Rank	Words	Appearance rate[%]					Sum of appearance rate
		Honda	Mazda	Toyota	Nissan	Subaru	
1	ご覧ください(Please look)	12.437	21.988	4.560	8.586	9.254	56.824
2	開催(Holding)	13.015	12.952	3.212	6.566	20.597	56.342
3	紹介(Introduction)	10.701	15.060	4.560	9.343	6.567	46.232
4	クルマ(Car)	4.411	15.964	8.601	5.556	8.060	42.591
5	詳細(Detail)	2.531	0.000	28.187	0.000	3.881	34.598
6	皆さま(Everyone)	2.820	3.614	0.000	2.020	18.507	26.962
7	お願い(Please)	4.121	9.337	0.000	3.030	6.866	23.355
8	行う(Perform)	11.424	5.422	2.487	3.157	0.000	22.490
9	いただく(*Polite Expressions in Japanese)	3.181	8.735	1.865	3.788	4.179	21.749
10	皆さん(Everyone)	6.797	0.000	2.487	0.000	6.567	15.851
11	発売(Release)	0.000	7.831	2.591	0.000	5.373	15.795
12	お待ち(Waiting)	3.471	4.819	0.000	0.000	5.970	14.260
13	応援(Cheering)	0.000	4.518	2.487	3.157	3.881	14.042
14	投稿(Posting)	2.603	7.229	3.005	0.000	0.000	12.837
15	東京モーターショー2019(Tokyo Motor Show 2019)	3.471	6.024	3.109	0.000	0.000	12.604
16	登場(Appearance)	0.000	3.916	4.870	2.525	0.000	11.311
17	写真(Photo)	3.688	3.614	0.000	0.000	3.881	11.183
18	良い(Good)	2.892	6.024	2.073	0.000	0.000	10.989
19	ドライブ(Drive)	0.000	4.819	2.176	3.283	0.000	10.278
20	目指す(Aim)	3.615	4.217	2.280	0.000	0.000	10.112

5.2 The relationship between tweet content and RT

In the previous section, the analysis focused on the words in the content of the tweets. This section focused on the relationship between the words in a tweet and the number of retweets.

To confirm the relationship between the numbers of retweets and word frequency, multiple regression analysis was used (Tanaka et al. 1983). The target data was set as the top 30 RT for each company. The objective variable was set to RT while the explanatory variable was set to the top 20 words of the word frequencies found in the previous section. Table 6 shows the results of R-squared in multiple regression analysis. As a result, the values of R-squared were all low and less than 0.4. This indicates the degree to which the model cannot be explained in multiple regression analysis. In contrast, there were the differences between companies, changing the type of variable would reveal differences. Thus, frequent words did not allow us to examine the influence of RT. On the contrary, examining characteristic words and groups of similar words provided clues to RT were found.

These results suggest that word groupings and specific proper nouns would be useful for examining the relationship between RT and tweet content.

Table 6. The results of R-squared in multiple regression analysis

Companies	R-squared
Mazda	0.387
Honda	0.258
Toyota	0.193
Nissan	0.135
Subaru	0.235

6. Summary and future works

This study examined the relationship between RT and the related Twitter factors. Furthermore, the content of the tweet was broken down into words by TM and the frequency of the words was counted. Main findings are as follows:

< i) Relationship between RT and Character counts >
    It turned out that the number of characters is not directly related to RT. However, there is still the possibility that Twitter, with its character limit, has a connection to RT on whether it has more or less characters.

< ii) Relationship between the average of RT and the timing of a posting tweet >.
    Trends in the number of tweets were identified for each data set. In hours, the coincidence of the time when most tweets are tweeted and the average RT are comparatively high is productive tweeting. This is one factor that indicates that not only the content of the tweet but also the timing of the post is important. Additionally, strong positive correlations were also partially identified in the Top30 RT. This at least indicates that post timing is also important in getting a lot of RT

< iii) Relationship between the words used in many of the current tweets and the number of RT?>
    Frequent words did not allow us to examine the influence of RT. On the contrary, examining characteristic words and groups of similar words provided clues to RT were found. Moreover, these results suggested that word groupings and specific proper nouns would be useful for examining the relationship between RT and tweet content.

    In order to prove the above possibilities and construct a model RT and explore factors that increase RT for the purpose of this study, the following Future works should be considered.
- Instead of LIKE, this study should find a variable that predicts the number of RTs from the information at the time of tweet generation.
- Verify whether the proposal applies to other industries, including data sets from other industries.

Acknowledgements

References

Cohen, A. M. and Hersh, W. R., A survey of current work in biomedical text mining, *Briefings in Bioinformatics*, vol.6, no.1, pp.57-71, 2005.

Dentsu, 2018 Advertising costs in Japan, Available: http://www.dentsu.co.jp/news/sp/release/2019/0228-009767.html, October 12, 2020 (in Japanese).

Han, S., Min, J., and Lee, H., Antecedents of social presence and gratification of social connection needs in SNS: A study of Twitter users and their mobile and non-mobile usage, *International Journal of Information Management*, vol.35, no.4, pp.459-471, 2015.

Kitano, Y., Yamada, T., and Tan, K. H., A Twitter analysis of Japanese companies with text mining, *Reprints of The Society of Plant Engineers Japan*, 2020 Spring meeting, pp.113-118, 2020.

Ma, J., Tse, Y. K., Wang, X., and Zhang, M., Examining customer perception and behavior through social media research: An empirical study of the United Airlines overbooking crisis, *Transportation Research Part E*, vol.127, pp.192-205, 2019.

Ministry of Internal Affairs and Communications, Survey on time and information behavior of information and communication media, Available: https://www.soumu.go.jp/iicp/research/results/media_usage-time.html, April 28, 2020 (in Japanese).

Nakamura, Y. Takabayashi, N. Ohba, M. Yamamoto, H. and Murayama, H., Analyzing companies and industries' indicator by using text-mining ~The case of 2010 and 2012's environmental reports, *Oukan*, vol.9, no.2, pp.95-103, 2015. (in Japanese)

Omuka, I., The history of SNS, *The Institute of Electronics, Information and Communication Engineers (IEICE),* vol. 9, no. 2, pp. 70-75, 2015 (in Japanese).

Tse, Y. K. Zhang, M. Doherty, B. Chappell, P., and Garnett, P., Insight from the horsemeat scandal exploring the consumers' opinion of Tweets toward Tesco, *Industrial Management & Data Systems*, vol. 116, no. 6, pp. 1178-1200, 2016.

Twitter Help Center, Retweet FAQs, Available: https://help.twitter.com/en/using-twitter/retweet-faqs, August 17 2019. (in Japanese)

Tanaka, Y. and Wakimoto, M., *Methods of Multivariate Statistical Analysis.* Gendai-Sugakusha, 1983 (in Japanese).

TOYOKEIZAI ONLINE, Ranking of the top 300 companies with the most "advertising costs", Available: https://toyokeizai.net/articles/-/187757, October 29, 2020 (in Japanese).