Escola Tècnica Superior d'Enginyeria Informàtica
Universitat Politècnica de València

Coverage of social incidents in the Republic of Korea by the local online press media

Trabajo Fin de Máster

Máster Universitario en Gestión de la Información

Autor: Lee Jonghyuk

Tutor: Enrique Orduña Malea

Directora Experimental: Cristina I. Font Julián

2018-2019

Resumen

Las noticias online se han convertido en un elemento importante que pueden condicionar la percepción de la realidad de los ciudadanos. Corea constituye un caso de estudio especial en este sentido. Este trabajo parte de la hipótesis de que algunos agregadores de noticias pueden estar manipulando la exposición de ciertas noticias cubriendo determinados eventos, que este trabajo pretende verificar de forma empírica. Naver se utilizará como caso de estudio. El objetivo de este trabajo es pues comprobar qué noticias aparecen más frecuentemente en la web de Naver y verificar si existe alguna correlación entre la actividad en Twitter de congresistas nacionales coreanos. Para llevar a cabo este objetivo se ha creado un sistema de información para recopilar y analizar datos web publicados en Naver (via web scraping) y Tweets publicados en Twitter (via API). Tras ello se han aplicado distintos análisis estadísticos (clustering, clasificación, correlación). Los resultados indican que algunas noticias de algunos medios específicos aparecen más frecuentemente que otros de una forma poco natural. No obstante, no se ha obtenido una correlación elevada ni ninguna evidencia de causalidad entre estas noticias y la actividad en Twitter de los congresistas coreanos. En todo caso, los datos tampoco rechazan esta hipótesis. Se concluye que Naver proporciona una exposición a ciertas noticias y periodistas, aunque se precisa más datos y análisis para concluir la posible relación entre estas noticias y la actividad específica de los congresistas para generar tráfico y visibilidad sobre estas noticias.

Palabras clave: noticias, twitter, correlación basada en el tiempo, aprendizaje automático, clustering, clasificación..

Abstract

The online news are important and can influence citizens' perceptions of reality. Korea constitutes a special case of study in this regard. There is a reasonable hypothesis that some news aggregators can be manipulating the exposure of certain news covering certain events. We are going to test it empirically. Using Naver as a case study, our objective is to figure out which news from which press appears more frequently on Naver's website and to check if there is a correlation with activity in Twitter from National Assemblymen of Korea. To accomplish our goal, we have created an information system to collect and analyze web data both news from Naver (scraping) and Tweets from Twitter (APIs). Then we have applied some statistical analysis (clustering, classification, correlations, etc.) on these data. Some news from specific press do appear more than others, and this is not natural. However, the correlation of the news with the tweets cannot surely confirm causality (activity in Twitter from Assemblymen is influencing more traffic to some news). But data does not refuse this as well. Naver provided more exposure to some press and journalists. However, the impact of these news cannot be stated as an effect of Twitter use of Assemblymen. In conclusion, more data is needed to get more insightful findings.

Keywords : news, twitter, time based correlation, machine learning, clustering, classification.

Contents Table

Index of Tables

Index of Figures

1. Introduction

1.1. Background

In 2016, since the presidential impeachment started [27], a number of press, like KBS, MBC, SBS, JTBC etcetera, have been pouring the huge amounts of news every hour every day. Considering that there have been lots of the serious social issues, like a 2018 North Korea–United States summit [28], this phenomenon might be spontaneous. But in terms of the general public citizen reading the news, the amount of information is too many to read and understand.

General people have to live, that is, they have to work, eat and sleep. Therefore it is very difficult to read a quantity of the news every day [31]. In addition, it is necessary to remember all history of each news to understand it correctly. In fact, it is quite impossible to digest a bunch of the news as a normal person.

As of April 2018, 75% of Koreans are using Naver's news collection [1] [30], and Naver was in doubt about manipulating the news and its comments written by the public. [26] [29] Therefore, it is interesting to analyze how the news are arranged on a webpage.

Nowadays modern people have come to the point where they need a new technical system to judge the authenticity of information source and to find a relation between information in order to make a critical thinking. This demand is even more important because the power of the wakened citizen is absolutely necessary in order for the Republic of Korea to become an advanced country in the circumstance being changed suddenly of the Northeast Asia surrounding Korean [28]. Based on this, the development of a news analysis system was selected as the subject of this TFM to help citizens make right decisions.

1.2. News, Layout and Twitter

1.2.1. Screen layout of the news

Now that paper newspapers have been decreased [32], most of the news is being broadcast to the general public through the Internet.

In the newspaper's days, the headline was placed on the first page of the newspaper. Nowadays it is still valid even on the web news. Therefore, screen layout of the news is very important factor to recognize the intention of each press.

In this study, it will be analyzed editing behavior of Naver about the news screen layout. Naver is the most representative portal of Korea. Naver has been explaining that he is not a journalist himself [1], but actually he acts like it is even though it is legally an internet portal service provider. Naver arranges he news published on its webpages arbitrarily and selectively. This behavior is no different from what the news editorial department does in a normal press company.

It is said in a statistics where 75% of koreans are watching the news on Naver's pages time to time. [1] [30] This means that there could be a dangerous possibility if Naver manipulates the layout of the news.

In addition, some public opinion, a handful of media outlets and political parties have raised suspicions about Naver's alleged manipulation of news [1] [30]. Therefore, analyzing how Naver does the news screen layout is a very meaningful study, and it could be an opportunity to find out the credibility of the suspicion raised recently.

1.2.2. News and the correlation with Twitter

There have been being constant suspicions that a core group of political peasants manipulates the public opinion in cooperation with conservative politicians, conservative media, conservative civic groups, and conservative large companies. Although this is a very interesting phenomenon, it could also occur a social disorder so that it is needed to know or to be aware of if those things are true or not.

Therefore, it would be interesting to figure it out whether there is a political cooperation, which means that the conservative media and the conservative politicians cooperate on who transmits the news, when the events happen, how to manipulate the public opinion at the same time.

The theme of the news means the event. What a news delivers to the public by the media, politicians, etcetera, is about an event. This idea can be also applied to the tweet. Hence, whatever it is among them, since they could have an event, they could be connected in any way. Therefore, it is important to clarify the meaning of the event that will be used through out the whole study.

Those words like "event" or "subject" may seem more natural in the area of the news and the twitter. But some technical words like "classification prediction" or "predicted value" will be also used in the same meaning of the event. This is obvious

because every news and every tweets will be marked as a number, which means what event it is about and which means the predicted value at the same time.

In the following sections we will use the general concept of classification predictions when discussing clusters and classifiers. However, after this chapter will discuss specific areas of news so that it will be used the following words in the same meaning. This is because dealing with social issues, the expression "event" or "subject" is more natural in the area of the news and the twitter, although the expression "classification prediction" is more accurate in terms of programming. Therefore, it will be used the words classified prediction, event, and topic in the future to suit the situation.

From now on, we will analyze the actual news and twitter by finding optimized combinations of clusters and classifiers. The reason why we dealt with the classification result of the combination of the cluster and the classifier and the classification result according to the combination is directly related to the question of whether the application result is reliable when applied to the actual situation. Therefore, we will analyze the news collected for a certain period of time with the combination of optimized cluster- classifiers obtained through the above-mentioned analysis process.

1.3. The analyzing target data and the purposes

1.3.1. Naver Mobile Homepage & News Service Main Page

Naver mobile homepage is the most visited page so that it is valuable to study patterns, which is about how to arrange the news on a screen layout. And secondly most visited web page is News service main page. Because every link of the news on the homepage is redirected to this area where lots of news are updated in real time and also classified into such as politics, economy, society, technology, world sports, entertainment and so on.

1.3.2. Politicians' Twitter

Twitter's analysis methods are diverse. The relationships between someone's followers and followings can be analyzed by analyzing the most often frequency of words on his tweets, or occasionally by reporting on what is happening in the news.

However, the purpose of this study is different. The main collective target of tweets is politicians, not ordinary people. There are lots of politicians through the time so that it is necessary to limit among them as a group of the current members of the National Assembly. The reason for this is that, as mentioned above, the relationship between press and politics is needed to be examined. Normally the current assembly men have an influence to the public and have a power to handle a variety of the social issues.

It is to know what and when politicians in a party are tweeting about.

1.3.3. Korea National Assemblyman Information

In order to analyze current members' tweets, they collect their list, affiliation party, and twitter account information.

1.4. Cluster & Classifier

This is the basic and inevitable knowledge to research all through out the study.

The absolute reason for the need for non-human machines in the big data field is that the amount of data is too large for human to interpret, as the name Big Data itself tells us. For example, the total number of daily news publications on a particular day is 1338 (This number came out from the collected news data during a month). No one reads 1338 news in a day. It may be possible if you decide and do nothing for the whole day and only read the news. But it is a waste of time. But what if we had to do this every day? It's like wasting your whole life. That's why we humans use machines. At least 10 people may be needed if people read and categorize 1338 daily news items, which is the number that was calculated by collected data from Naver's news web pages in this study. However, this researcher is one person and has other things to do.

However, luckily, the technology of machine learning [15] has been developed, and the technology of clusters required for this research has developed considerably. In

particular, thanks to cluster technology that has been successful in image recognition, we have been able to apply this technology to a specific area of news written in human language.

A cluster is a group of words. Humans have a mechanism for thinking through the experience or naturally tying similar groups together. This ability is a special ability to make human beings human compared to other animals, and often has a large error range in many cases. In particular, experience shows that the larger the number of objects that a person has to recognize, the larger the error range. Because it reaches the limit of memory. (I will not mention the unconscious world in this study.) Humans can not memorize large amounts of data at all, nor can they process them simultaneously. Therefore, it is important that the machine automatically analyzes patterns of various data and classifies them automatically.

By the way, the most important and key technology in this study is the clustering technology. If so, it is necessary to know how accurate the clustering technology is currently developed. Numerically the basic clustering technology is known to be somehow accurate. However, in some cases of datas such as a text and a picture except for numbers, it is known that the accuracy is questionable as it can be found in real life. Since this study applies the clustering technology to the news articles written in Korean, it is inevitable to confirm these doubts.

1.5. Hypothesis

The hypotheses presented in this study are due to the insight of the researcher who was born naturally as he / she actually encountered various news programs of various media in everyday life. It is true that insight as such is meaningful in itself, but it is somewhat less convincing for those who live in the 21st century to depend on human insight alone for all social phenomena. We believe that human insight is a great ability to measure, but we need a numerical basis to increase the confidence in the insight as well as the insights of what is measurable.

Therefore, it is necessary to verify the reliability of the insights on the hypotheses that the researcher has estimated using data analysis technology.

1.6. Secondary goals

Before getting into the 2nd chapter, it is necessary to confirm secondary goals in this study.

- To find out how Naver arranges the news on their homepage and on their news service pages.
- To find out what kinds of correlations are there between the press and the politicians.
- How a bunch of datas related to the news can be analyzed programmatically with a high reliability.

2. Methodology

The study tries to approach to in views of three perspectives. First, we will explore how Naver could try to manipulate the public's opinion by placing the news on the web pages. Second, we will analyze what news is talking about and how we can find a correlation with the tweets. It requires machine learning skills. Therefore, we will study the clusters and the classifiers. Third, we will analyze the subject of the news, and examine the correlations between the politician and the media and then will derive the hypothesis of a conspiracy about the recent social issues.

2.1. Conceptual overview of Data Analysis

2.1.1. Web data collection

2.1.1.1. Definition of collecting

Data collection refers to two types of actions: creating data on its own or importing data from outside. In many cases, developers make the mistake of parsing and storing raw data at the same time as they are collected.

However, collecting and storing all the information that Raw Data has are advantageous for both system development and data analysis. First, it is due to the system of denial of service from the service encountered during the web scraping process. Some web service companies have various forms of defensive system. In this case, data collection and parsing must be developed separately because data loss may occur during the data collection process.

Second, repeated redesign and redevelopment during data analysis are inevitable. If data structure is determined at the beginning of development and "collection → parsing" process is combined into one, there is a high possibility that data that may be needed later may be lost during the collection process. Therefore, the definition of collection in this project only includes collecting and storing Raw Data itself.

2.1.1.2. Collection steps

The collection methods can be broadly classified into three types.

- The first is to utilize the open API, legally and officially provided by the service provider.
- The second is a way to scrap webpages that are legitimate but informally not provided by service providers like open APIs.
- Third, there is a way to illegally hack. However, this method is not discussed in this study.

2.1.1.3. Collecting target and related technology

In the table <1> below, the brief summary is shown, what the source targets are, how to collect data according to each of the source targets and their specific collecting technologies including the API, Webscraping.

Target	Method	Collecting tech.
Naver News Collection Service	Open API	• Naver Open API (news search API)
	Webscraping	• Requests: HTTP for Humans • Pandas.read_html
Twitter of journalists / politicians	Open API	• Twitter API • Tweepy
Web services of each press	Webscraping	• Requests: HTTP for Humans
Korea National Assemblyman Information	Webscraping	• Requests: HTTP for Humans

Table <1> Collecting method & tech.

In the following chapters, more information about the technologies of each target will be explained in details.

2.1.1.3.1. Naver News Search Open API

Naver, an Internet portal service provider, offers its diverse data and services as open APIs. Among them, there were some APIs to be used in this study, but they did not get satisfactory results, so I used only one API. The screenshot below is the first screen of the API used in this study. Unfortunately, the service only supports Korean.

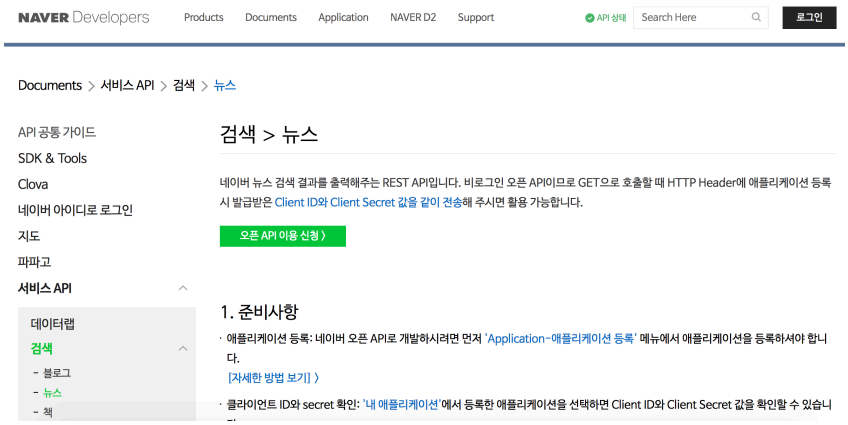


Fig.<1> NAVER Developers site where News search OPEN API [3]

2.1.1.3.2. Pandas.read_html

Pandas is a data analysis package for Python. The “read_html” function is very useful when scrapping web pages. This is easily solved by the Pandas function because you have to parse the table information in a page programmatically unless you do not use it. [4]

2.1.1.3.3. Requests: HTTP for Humans

Python also provides urllib, but it is not specific to web scraping. However, the Python Requests package is specialized for web scraping. [5]

2.1.1.3.4. Twitter Open API

It provides all the APIs of Twitter. In particular, in order to use Tweepy, the user must go through the developer registration process to obtain the authentication key and the access key. [6]

2.1.1.3.5. Tweepy

You can use the platform provided by Twitter directly, but as mentioned on the homepage as the core article, the Tweepy package provides an interface that is easier to use [7] programmatically. In many books on data analysis, you can see that this package is used for example implementations.

Fig.<2> Tweepy Homepage

2.1.2. Analysis

2.1.2.1. Definition of analyzing

Analysis is a broad activity that organically reinterprets source data from a variety of sources, both collected and internally generated. For this reason, the term : data analysis, is often used in combination with the term : data science. In this study, the concept of analysis is intended to be used by reducing the concept of the general concept. Therefore, the term is necessary to be abbreviated as a term ‘analysis’ rather than ‘data analysis’.

According to the new term of ‘analysis’ in this study, this method can be split into two steps, which are the parsing and the analyzing.

2.1.2.2. Parsing the source data and related technology

Collected data from various types of sources must undergo basic parsing procedures. This is the same as the data preparation in the general data analysis. The parsing of the original data is to parse the raw data, which not only parses the textual data encoded in Unicode, Bytecode, UTF-8, but also determines and interprets the data type for each data according to the data structure definition.

The types of data parsing can be roughly divided into two types as follows.

2.1.2.2.1. Parsing API Responses

The easiest way to do this is with a small number of service providers. Most of the API's response is made up of JSON, but since it is raw data, it must be converted to a real JSON data type.

2.1.2.2.2. Parsing web pages

It is a much more difficult form of parsing the API's response than it is to collect data in an unofficial way. Each service provider needs to have a special module that can parse the HTML DOM into your way. Beautiful Soup is used in this study, which is a package for Python. [8]

Target	Parsing tech.
Naver News Collection Service	• Beautiful Soup • JSON
ETRI language analysis service	• JSON
Twitter of journalists / politicians	• JSON
Web services of each press	• Beautiful Soup

Korea National Assemblyman Information	• Beautiful Soup

Table <2> Parsing Raw Data by Collection targets

2.1.2.3. Analyzing the parsed data and related technology

In the next few steps, more information about the technologies being used through the whole project.

2.1.2.3.1. NLTK

It stands for Natural Language Toolkit, a package for Python for natural language analysis. [9] Despite its support for various languages, unfortunately, Korean is not enough supported. It provides various analysis libraries mainly for roman-based languages.

2.1.2.3.2. KoNLPy

It is an abbreviation of Natural language processing of the Korean language. [10] It is a package for Python which is made by Korean graduate students. The existing NLTK also supports Korean, but its function is not enough. In the paper,

> *"Beautiful, but somewhat complicated, language is the 13th most used language in the world. In the meantime, Korean morphemes have been developed to extract useful features from complex subtle Korean texts. KoNLPy does not want to create another tool with the same functionality. Rather, it builds a layer above the existing tools to look further."* [11]

2.1.2.3.3. Public artificial intelligence open API · DATA service portal

It is an AI platform created by the government of the Republic of Korea [12] and cooperated with national affiliate ETRI (Electronics and Telecommunications Research Institute) [13], mainly focusing on interactive artificial intelligence analysis such as language analysis and speech recognition. [14].

In addition to the KoNLPy mentioned above, there are Korean analysis libraries of this kind. The following tables are representative libraries.

이름	연도	언어	라이센스
KTS [14]	1995	C/C++	GPL v2
한나눔 [13]	1999	Java	GPL v3
MACH [11]	2002	C/C++	custom
Arirang	2009	Java	Apache v2
꼬꼬마 [7]	2010	Java	GPL v2
KoNLP [5]	2011	R	GPL v3
MeCab-ko [6]	2013	C/C++	GPL v2, LGPL, BSD
KOMORAN	2013	Java	custom

Table <3> Several open-source Korean stemmers and development languages [11]

As a result of comparison doing the performance tests on a process of this study between KoNLPy and ETRI AI Language Analysis API, finally it is the best option to select ETRI API because it is faster and more accurate.

2.1.2.3.4. Scikit-learn

It is the best-built package for Python in the field of machine learning. [15] This is the core library of this paper and provides various algorithms for various types of big data about clustering technology for unsupervised machine learning and classifying technology for supervised machine learning.

2.1.2.3.5. Pandas

It is the best package for data analysis and is most widely used with Numpy and SciPy. [16] As the most fundamental library in this research, it is possible to handle all data types of numeric and character types. The handling of data frames is much easier to handle than the package for handling data frames in Node.js, and the documentation and examples are abundant. I originally started with Node.js, but because the manipulation of the data frame was more difficult than Pandas, I re-studied Python and turned to using this library.

2.1.3. Report

2.1.3.1. Definition of reporting

The definition of the report referred to in this study is a broad concept that includes reporting for reviewers and general readers of the research papers, and reporting for those who design, develop and administrate the entire system. The report for the first subject is a general concept of the report and it will show the final result of the research using tools such as charts and graphs. The report for the second target is a report on all the occasional occurrences throughout the whole data analysis, and it is necessary to develop the system activity status, to measure the execution time of the analysis activity, to do the recursive redesign according to the agile development methodology and to obtain the various, necessary datas for redevelopment Tool.

2.1.3.2. Report steps

2.1.3.2.1. External reporting

It is a report to the general public, which means that the final result is shown in the form of a chart or graphic. All tables and figures inserted in this paper correspond to this form. Based on the final result data, it dynamically represents the core idea through clients such as web pages or Jupiter Notebook.

2.1.3.2.2. Internal reporting

It is mainly based on the logs, which is shown in the command line interface through the any kind of terminals of whatever computer is. It is useful to inform the researchers themselves or developers of the status of data analysis operations. In this study, two tools are used like the terminal of Apple Mac, Jupyter Notebook [18].

On the other hand, the administration's internal web page can be used. It is a good tool after the completion of the project to maintain the systems. However, as it is mentioned before, this is not necessary in this study because this study has been in the process of the project. Many modules in this system have to be changed and re-invented. Because of this reason, the administration's internal web page was dismissed.

Back to the terminal tool, this is very useful because it shows lots of information of the program running to watch the collection status, the status of the collection completion, the status of the completion of the parsing, the execution time of the clustering, the execution time of the classifier, the best optimized parameters and so on.

2.1.3.3. Reporting subject and related technology

2.1.3.3.1. Matplotlib

It is a package for Python that is similar to MATLAB and is designed to visualize data. [17] Numerous types of visualizations are possible from basic 2D shapes such as Line, Box, Bar, Histogram, Pie, Scatter, and Radar char to 3D. The final report of this study was selectively used to match the nature of the core subject.

2.1.3.3.2. Jupyter Notebook

It is a development tool mainly used by Python beginners. [18] I usually do not use this development tool well, but I used it to identify the final results to insert into the paper. With the inline command for more efficient use of the Matplotlib package, you can quickly see a large number of visualization results in Jupiter Notebook. It also provides publishing capabilities for HTML file types for presentations, making it convenient for external reporting.

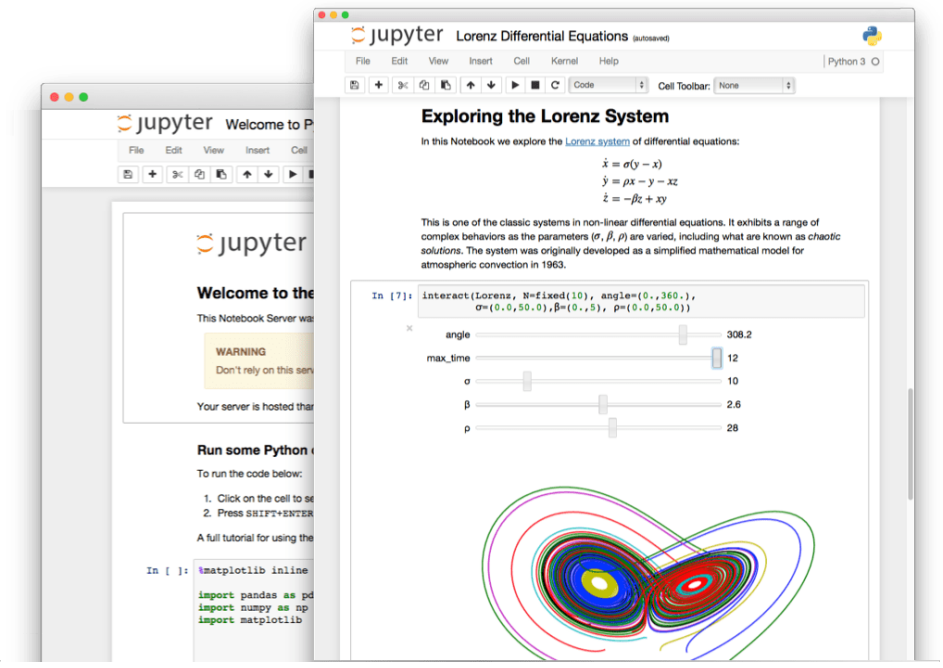


Fig. <3> Example of using Jupyter Notebook [18]

2.2. System design

In this section, It will be explained in detail how to conduct this research, that is, the overall project. The figure <4> will briefly show how to build a system for project execution, how to collect data, how to analyze it and how to express it in a more technical viewpoint about each targets corresponding to the three stages of data analysis mentioned in the previous introduction chapter.

2.2.1. System architecture

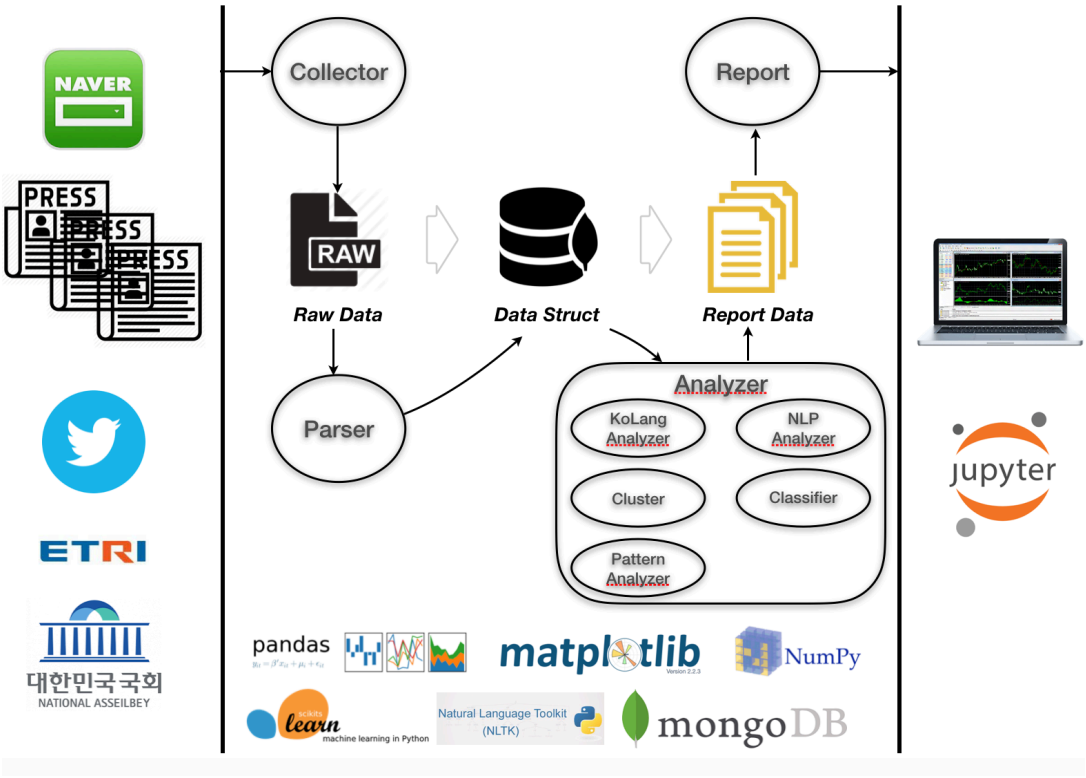


Fig. <4> System Architecture

The figure <4> shows a schematic diagram of the overall system of this project. On the left are the main collection objects, on the right are two interfaces for reporting. The center is the internal structure of the system. It is broadly divided into four areas: Collector, Parser, Analyzer, and Reporter. I expressed this in four steps being added the Parser, unlike I mentioned above that there are three steps. The reason is to distinguish between raw data and data structure. The separation of these two is inevitable because of the large amount of data to be analyzed. If you combine these two types of data or miss one, then you will be in great trouble in a process of the analysis later.

The most important part of the internal system architecture is also the analyzer's area. Its five most important tasks are described. The data analysis open libraries required to carry out this research are shown below.

2.3. Data structure design

Since the sources of data to be collected are various, it is necessary to interpret the data structure for each data. In addition, a common data structure is needed internally in the system because the collected data must be analyzed in an integrated manner.

In the next few steps, it will be shown that some core data structures are used in this study. Even though there are lots of mongo database's tables, it is not necessary to know every tables' data structures. Below is the core data structures.

2.3.1. Screen layout data structure design

This the data structure of any screen layout being collected from any web pages. A page's layout can be interpreted a group of blocks or areas, which are the higher level of HTML tags in the body tag. As you can see below in the table <4>, category is divided into 3 components, Target, Position and Time. Screen layout has just a target(news' brief information) information and a position in a web page. But in this study, it is necessary to calculate the collecting time from a web page because any web page is changed instantly depending on service providers' policies.

Category	Column name	Data type	Definition
Target	href	String	URL of the web page to be collected
	new_title	String	Title of the news posted at each location
Position	page_name	String	The name of the web page to collect
	block_name	String	Large block of pages
	block_pstn_rank	int	Position's ranking of a block in-page (Top-down)
	news_pstn_rank	int	Position's ranking of each news in-block (Top-down)
	first_page_shown_TF	Boolean	Whether positions is displayed on a small mobile screen when the web page is first loaded
Time	collect_dt	datetime	Date and time (in milliseconds) when the web page is collected.

Table <4> Data structure of Layout for every websites

2.3.2. News data structure design

Second, this data structure design is very important as much as the first data structure of Screen layout. This handles a news or an article itself. It is possible to obtain a link from the screen layout easily, and then every articles from collected information previous in Screen layout should be collected one by one. In this data structure, other information is more important to analyze the entire this project like WHO, WHEN, WHAT, HOW and so on. The data structures of Screen layout and of News are connected by the position and the target information like the layout id and news' url. By parsing and analyzing contents of each articles, it can easily be got the rest of core information.

Category	Column name	Data type	Definition
Position	lay_id	ObjectId	MongoDB ID of screen layout table
Target	news_url	String	URL of the corresponding web page of the news itself rather than the URL of the media web page
WHO	press_com_nm	String	-
	jounalist_nm	String	-
WHEN	collect_dt	Datetime	Date and time the news was collected
	publish_dt	Datetime	Date and time the news was first published
	modify_dt	Datetime	Date and time the news was last modified
WHAT	naver_event_cate	String	Event category name arbitrarily classified by Naver
	event_id	Int	Indicate what event the news deals with. Integer values classified by future classifiers
HOW	title	String	News title
	body	String	News body
ETC.	like	Int	The average user can click on the news, article likes, cumulative

Table <5> Data structure of News for every web-published news

2.3.3. Cluster data structure design

To analyze a bunch of datas after collecting & parsing them from Screen layout information and News, it is mandatory to use Machine learning because it is impossible to read every web pages' layout and each news from them as a human being. Therefore, after applying the cluster technology of machine learning on both side, which are the information from screen layout and news, analyzed datas should be saved like the table <6>.

The category column shows which datas or which tables in the Mongodb are used for the clustering, which clustering methods are used for them and what are the results of the clustering. There are more detailed descriptions on the definition column. (As the philosophy of Software Engineering, column names are written in the programming way.)

Category	Column name	Data type	Definition
Target	clst_tbl	String	Target table name to be clustered
	clst_col	String	The name of the column to be clustered
Method	sampling	Int	Number of samplings to cluster
	n_clusters	Int	Number of cluster groups
	algorithm	String	Cluster algorithm
Result	KeyidLabel_dicli	dictionary-list	Cluster result. A list of target IDs and event IDs
	exc_time_sec	Datetime	Clustering execution time (seconds)

Table <6> Data structure of Clustered datas

2.3.4. Classifier data structure design

After the application of the clustering was finished, daily updated data should be analyzed again by using the results of the clustering.

In this table <7>, 3 components in the category column are important, Target, Method and Result. Target means the new datas that has to be analyzed. Method means which clustering technology is used. Finally, Result is exactly what the researcher wants to get.

Category	Column name	Data type	Definition
Meta Info.	key_col	Sring	The column name to use as the key value of the document in the target table to be classified.
Target	test_tbl	String	Target table name to sort
	test_col	String	Target columns name to test
Method	clst_id	ObjectId	The corresponding document ID in the cluster table containing the information on which the training data required for classification was clustered.
	clf_algorithm	String	classifier algorithm
Result	KeyidLabel_dicl i	dictionary-list	Classification result. A list that combines the target ID and the Predicted value
	exc_time_sec	Datetime	Classification execution time (seconds)

Table <7> Data structure of Classified datas

2.4. Research topics and its method

In this section, it will briefly be explained how to proceed the analysis about each topics.

2.4.1. Screen layout of news

The figure <5> represents the first page of Naver Mobile homepage. Interestingly, as you can see below, Naver's first few layouts of the homepage are filled with lots of news even though it is the Internet portal service provider. (From the left to right, the first, the second and the third layout)

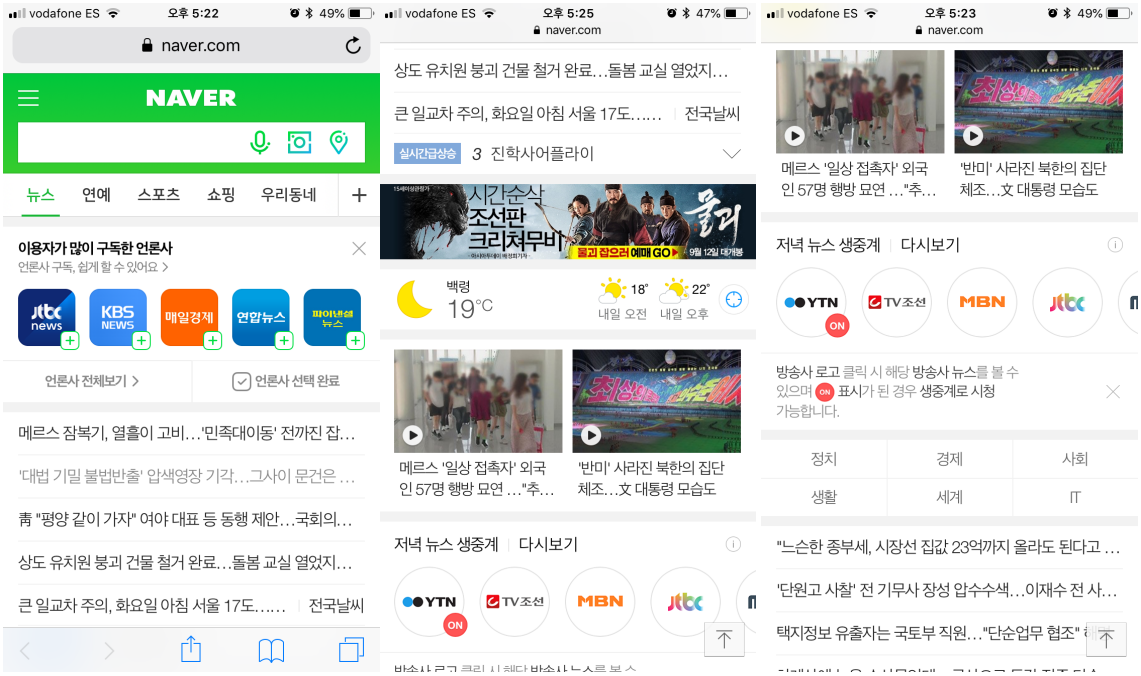


Fig. <5> Naver Mobile homepage first screen and next scrolled screen captures

This study mainly attempts to analyze various aspects of the news on the above screen.
First, it will be calculated how long a news is occupied at a particular location in the screen layout. In general, it could be considered that it is common that a news exposed at a specific location on the screen is selected randomly, or that events being currently issued are selected statistically. However, any common senses that we believed could turn out to mistakes and sometimes this phenomenon was proved historically. Maybe it could happen on this topic and that is why some recent suspicions by the public are necessary to be examined.

Naver's homepage in figure <4> could be hierarchically divided into some blocks(areas) as a way 'top-down' and each block has some positions of each news.

However, in order to analyze the data, these kinds of way have to be transformed into a programmatic thinking way.

This is the representation form used in the chapter 3 about how to specify the locations of each news in a web page in a programmatic way :

[Location index rule]
page name : in-screen area name : arrangement order of the area : arrangement order of news in the area : whether the first screen is exposed.

The splitter ':' will be used for dividing each metadata of the location.

2.4.2. Cluster & Classifier

To apply the clustering and the classification to the news and the tweets, they will be clustered and classified in several ways by using lots of conditional parameters.

For example, the titles and the body texts will be clustered. In addition to that, it can be applied by 2 types of algorithms and the varying number of clustered groups will be given.

In this study, the normal process is like this :

- Make several clusters by considering as many input conditions as possible.
- Obtain as many classifiers as possible by using several clusters and by changing the input conditions.
- Find a few of better classifiers after performing some statistical investigations on the predicted results.
- Choose the only one optimized classifier and apply it to the test data set.
- Repeat the process above while changing the combinations of the variables.

There are lots of the input conditions, which means those could be a parameter of algorithms or could be a train data set or could be a modified train data or could be a type of algorithms and so on.

2.4.3. News

Each news has these information : title, body text, published date, reporter name, press name and so on. Every collected news information are structured and saved in the MongoDB. The main datas are the title and the body. The rest of them will be used to calculate some statistical outcome.

The title, the body text will be clusters by scikit-learn python library and then they will be classified to specify what a news is about.

2.4.4. The correlations of News & Twitter

To solve one of the secondary goals, it is programmatically presented below in the figure <6>.

In the figure <6> the source code is a simple function that iterates 423 events and finds the correlation between news and tweets for each event. It is important to note that the source code lines 178 through 181, because the daily data of the news and tweets for the same event may not be present on either side. Therefore, we can calculate the right correlation coefficient by calculating the two data together by centering the daily data of the case. After some additional programmatic processing, the correlation coefficients shown in the following table can be obtained.

```
167 def CorrCoeff_Calculation_by_event(news_df, tw_df, intersections):
168     import Report
169 
170     corr_dicli = []
171     for event_num in intersections:
172         news_g = news_df[ news_df['predicted']==event_num ].groupby('입력일시').count().loc[:, ['news_id']]
173         tw_g = tw_df[ tw_df['predicted']==event_num ].groupby('created_at').count().loc[:,['tw_id']]
174 
175         news_data = Report.TimeResampled_data(df=news_g, rsp_period='24H', agg='sum')
176         tw_data = Report.TimeResampled_data(df=tw_g, rsp_period='24H', agg='sum')
177 
178         if len(news_data) >= len(tw_data):
179             cmb_data = news_data.join(tw_data)
180         else:
181             cmb_data = tw_data.join(news_data)
182         print(cmb_data)
183 
184         corr = cmb_data.corr()
185         corr = corr.fillna(0)
186         corr_val = {'01':corr.values[0][1], '10':corr.values[1][0]}
187         corr_dic['event_num'] = event_num
188         corr_dicli.append(corr_dic)
189 
190     return corr_dicli
```

Fig. <6> Calculation of event-related correlation of news and tweets

2.4.5. Politicians' information

Normally the National Assembly are 300 in Korea. But it was found that just 59 members of them use and activate their twitter accounts after investigating it manually. Because most of them are still known to use blogs more than tweets as a means of communicating with the general public. Hence, a total of 59 members' tweets are collected.

Some information on the National Assembly [26] will be gathered from the National Assembly web site of the Republic of Korea by using Beautiful Soup4 python library. It will be analyzed how the members of the party with their political tendencies have mentioned about an event. However, considering the fact that the number of members of the National Assembly is about 300, it could be better to

analyze the political tendencies of the political parties belonging to the members of the parliament, rather than to mention one member of the parliament directly.

In addition, the political spectrums of the political parties in Korea follow the common social sense of the public or the press because any political parties do not say that they are Right-wing politics or the others. These spectrums will be used when it is analyzed in the section 3.4.

2.4.6. Machine learning theory

There are lots of machine learning technologies in each programming language at the present. But this study uses Python so that scikit-learn library will be used for the analysis. Moreover, just two categories will be used, which are Clustering and Classification, except for the rest of scikit-learn's libraries.

The figure <7> shows how to choose a proper algorithm depending on his demands. It is easy to use because it seems to follow some steps of the questions. But sometimes, it does not work properly the final algorithms that it recommends us. Anyway, in order to do the researches as many as possible it is not quite bad idea to use the below guide-map.

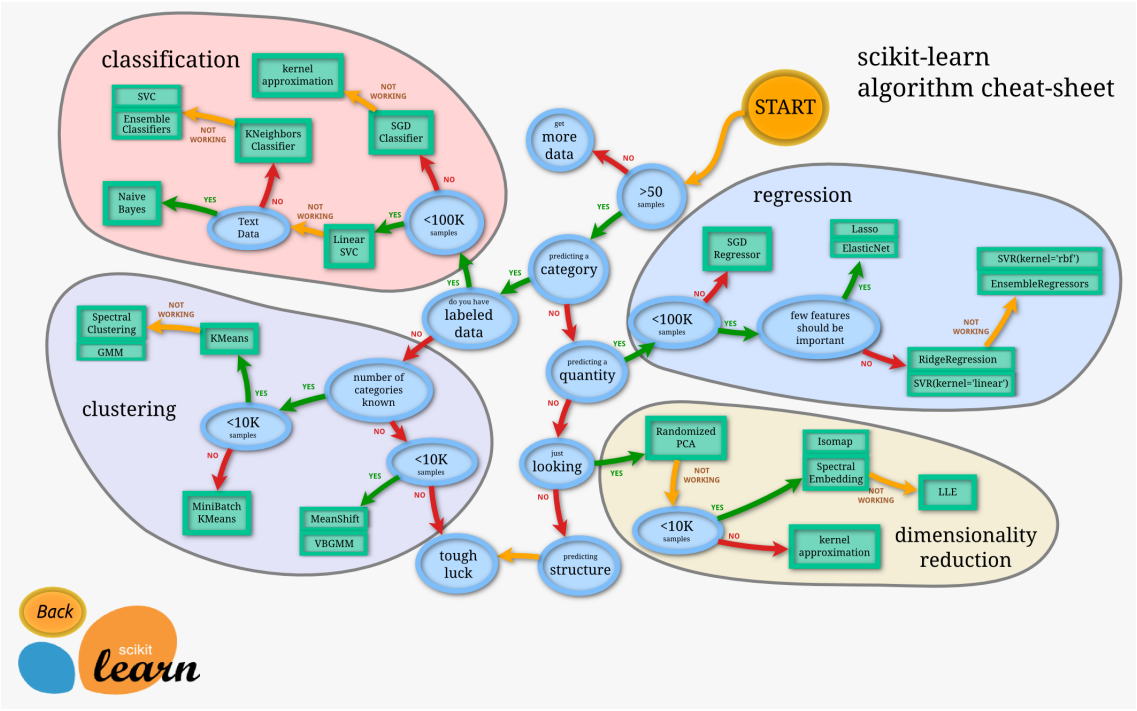


Fig. <7> Choosing the right estimator [19]

As it is said in the link of the figure <7>, it was so hard to choose which algorithms it should be use in this study. Actually, as a beginner in the machine learning field, it is impossible to know every theories, theorems and algorithms. Therefore, it was recommended to look for right algorithms by checking several conditions of this study. For example, its case was like below :

- More than 50 samples
- Predicting a category
- Do not have labeled data
- less than 10K samples

2.4.7. Cluster

Based on the right estimator guidance provided by scikit-learn and the conditions above, 4 algorithms, which are Mini Batch, K-Mean, MeanShift and VBGMM, are the candidates to try.

But among them, K-Means and Mini Batch K-Means are finally chosen in this study and the reason for this was that the outputs of two things were more reasonable than others after testing few sample datas.

2.4.7.1. Comparison of KMeans & MiniBatchKMeans

The key equation of this algorithm is easy to understand. The equation is as follows.

$$\sum_{i=0}^{n} \min_{\mu_j \in C}(||x_j - \mu_i||^2)$$

where,

n : the number of samples

sample : X

disjoint clusters : C

mean : μ_j

In short, it is the minimum value of the sum of the squares of the absolute values of the differences of the vectors. These concepts are easy to understand because they are already learned in high school mathematics. For more details on this algorithm, see the related links. [20]

The difference between the two algorithms to be used in this study can be seen in the corresponding links. As a result, the researcher tested the news data and found that the result difference between the two was much larger than the result difference of the link. I think this is due to the difference in performance between digit-based and text-based calculations, such as news.

Differences in the link :

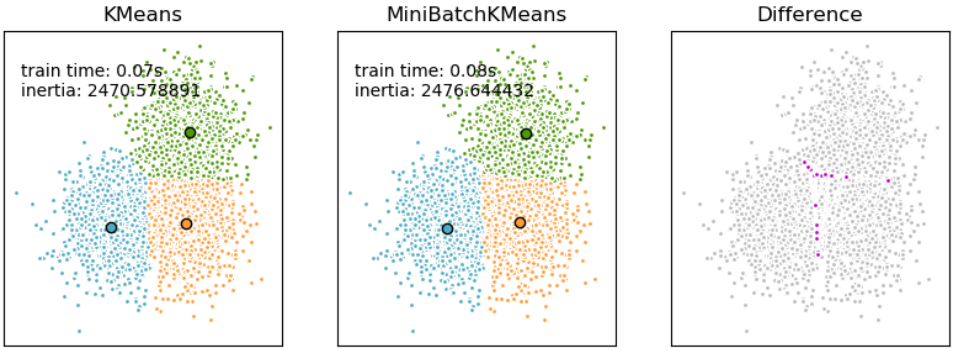


Fig. <8> Difference of the results between two algorithms [20]

Differences in this project using News :

where,

- Green = K-Means

- Red = Mini Batch K-Means

- label = Integer result of news from cluster

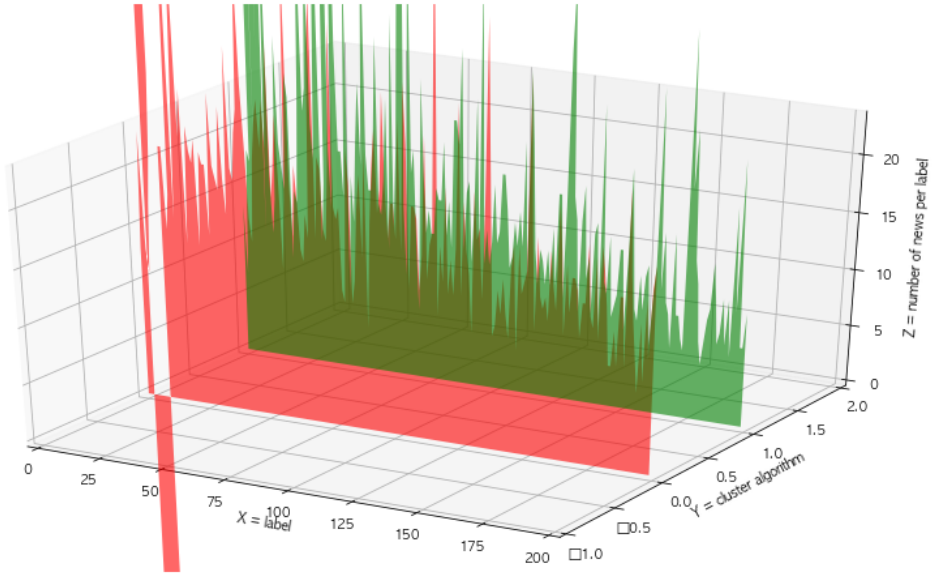


Fig. <9> Difference of the results between two algorithms about News

The shape of the graph is different, but the difference in performance can be confirmed. The X axis represents the label, which indicates the group number of the group. The distribution of green and red groups shows that there is a big difference at a glance.

2.4.8. Classifier

It is the normal definition of the equation of Perceptron. [21]

$$f(x) = \begin{cases} 1 & \text{if } \mathbf{w} \cdot \mathbf{x} + b > 0 \\ 0 & \text{otherwise} \end{cases} \qquad \sum_{i=1}^{m} w_i x_i$$

In the modern sense, the perceptron is an algorithm for learning a binary classifier: a function that maps its input x (a real-valued vector) to an output value (a single binary value) :

where,

w is a vector of real-valued weights

m : the number of inputs to the perceptron

b : the bias which shifts the decision boundary away from the origin and does not depend on any input value

The perceptron algorithm is a kind of classification algorithm in linear-model. It is used two classifier's algorithms in this study, which are Perceptron and MultinomialNB. Finally Perceptron won the prize between them even though both also are for analyzing text. Actually it is not easy to know why right now. This is a future homework to study further.

3. Results

Now we are going to examine the results from the whole study based on what is mentioned before. To get to the point, it is going to be described a kind of brief explanations about each research topics and their results.

3.1. Screen layout of news

3.1.1. The number of the news in each location

In chapter 2, it was already explained how to structure a web pages to analyze it.

As you can see in the table <8>, those positions are represented like in the ‘location’ column.

location	news title	location occupancy duration
Naver_mHome:character-type_news_list:1:1:Y	279	279
Naver_mHome:character-type_news_list:1:2:Y	336	336
Naver_mHome:character-type_news_list:1:3:Y	399	399
Naver_mHome:character-type_news_list:1:4:Y	466	466
Naver_mHome:character-type_news_list:1:5:Y	229	229
Naver_mHome:character-type_news_list:1:6:Y	236	236
Naver_mHome:character-type_news_list:1:7:Y	77	77
Naver_mHome:character-type_news_list:1:8:Y	17	17
Naver_mHome:character-type_news_list:1:9:Y	2	2
Naver_mHome:thumbnail_news:2:1:Y	356	356
Naver_mHome:thumbnail_news:2:2:Y	344	344

Table <8> The way to combine information of location in a webpage (unit: count)

On the right-side in the column of location occupancy duration, we can see that the number of news is different according to the left "location" index, which means that the exposure time of the news differs according to each location.

In particular, the numbers of the news being located in ‘character-type_new_list: 1:7~9’ are just a few comparing to ‘thumbnail_news’. This may be obvious because those are in the third screen, as you can see in the figure <4>. It means that Naver does not need to change a news quickly because it has a low possibility to be exposed to the users.

3.1.2. The occupancy time of each location of news on the first screen

3.1.2.1. Top ranked 15 news of the occupancy time

The next step is, in the table <9>, to look out a list of location occupancy times(secs) for each news on a specific location. In this case the most visible location of the general users is chosen, which is "Naver_mHome:character-type_news_list:1:1:Y" . The table <9> below shows only the top 15 ranked list of 279 news items occupying this particular location. Unexpectedly, just fews news had been occupying a specific but important location for a long time. If comparing the first ranked one to the 15th ranked one, it could be so easily recognized.

	news title	location occupancy duration	location
181	사설	2603914.503	Naver_mHome:character-type_news_list:1:1:Y
215	일문일답	2354164.547	Naver_mHome:character-type_news_list:1:1:Y
220	전문	1965810.089	Naver_mHome:character-type_news_list:1:1:Y
105	기상특보	574871.668	Naver_mHome:character-type_news_list:1:1:Y
242	태풍경로	84005.316	Naver_mHome:character-type_news_list:1:1:Y
11	"메르스 확진자, 동선 CCTV 영상 확보…밀접접촉자 22명"	41396.965	Naver_mHome:character-type_news_list:1:1:Y
182	사흘간 6차례·11시간…이산가족 상봉, 어떻게 진행되나	38101.452	Naver_mHome:character-type_news_list:1:1:Y
265	특보	35049.474	Naver_mHome:character-type_news_list:1:1:Y
23	'김경수 공모' 물증 제시 못한 특검 발표…재판 전망은?	30602.493	Naver_mHome:character-type_news_list:1:1:Y
281	피해 키운 '샌드위치 패널·유독가스'…스프링쿨러 작동도 의문	29701.960	Naver_mHome:character-type_news_list:1:1:Y
266	특사 파견으로 협상 교착 반전 모색…중재 역할 '시험대'	28378.477	Naver_mHome:character-type_news_list:1:1:Y
198	안희정 2심 판결은 달라질까…'미투 재판'에 미칠 영향	28201.468	Naver_mHome:character-type_news_list:1:1:Y
153	메르스 환자 3년만에 발생…밀접접촉 20명 격리조치	28072.060	Naver_mHome:character-type_news_list:1:1:Y
183	상봉 기쁨에 어느 때보다 벅찬 하루…향후 일정은?	27900.732	Naver_mHome:character-type_news_list:1:1:Y
260	특검, 김경수 지사 구속영장 청구…金 "강한 유감"	27601.059	Naver_mHome:character-type_news_list:1:1:Y

Table <9> Occupation time of Naver_mHome:character-type_news_list:1:1:Y (unit: secs)

3.1.2.2. The distribution of the occupancy time

When we look at the distribution of occupancy time of 279 news items, this unusual things return out more obvious. The figure <10> shows it.

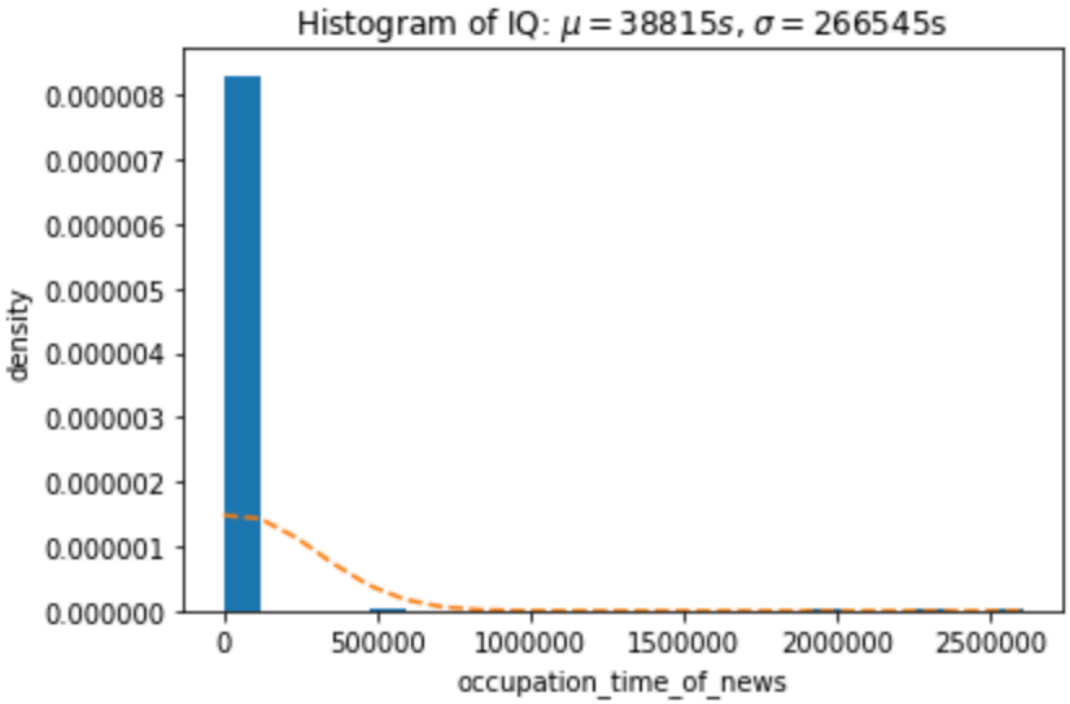


Fig. <10> Position occupation time distribution of 279 news occupying a specific position

It is the very unexpected histogram graph in a view of the proven theory that more than 1000 samples statistically make Gaussian distribution curve. It can be seen that the error range from the minimum of 240 seconds to the maximum of 2,603,914 seconds is large and that an extremely small number of news has location occupancy duration. This could be an evidence that Naver, proposed in the hypothesis, is manipulating the public opinion with the power to edit the news.

3.1.2.3. The composition ratio of the occupancy time

In addition, we could look at this suspicion in another way. That is, we could calculate the proportions of media companies relating to whether Naver could be manipulating the locations of the news by collusing with a certain media based on the validity of the hypothesis analyzed above.
Figure <11> is calculated about the figure <10> at a specific location (Naver_mHome:character-type_news_list:1:1:Y), and it shows how many the composition ratio just few medias have.

Yonhapnews is not the biggest media nor the influential media in Korea. New1 & newsis neither. Generally thinking, it is not logical that Naver put their news for a long time by using just 3 not-famous media's news.

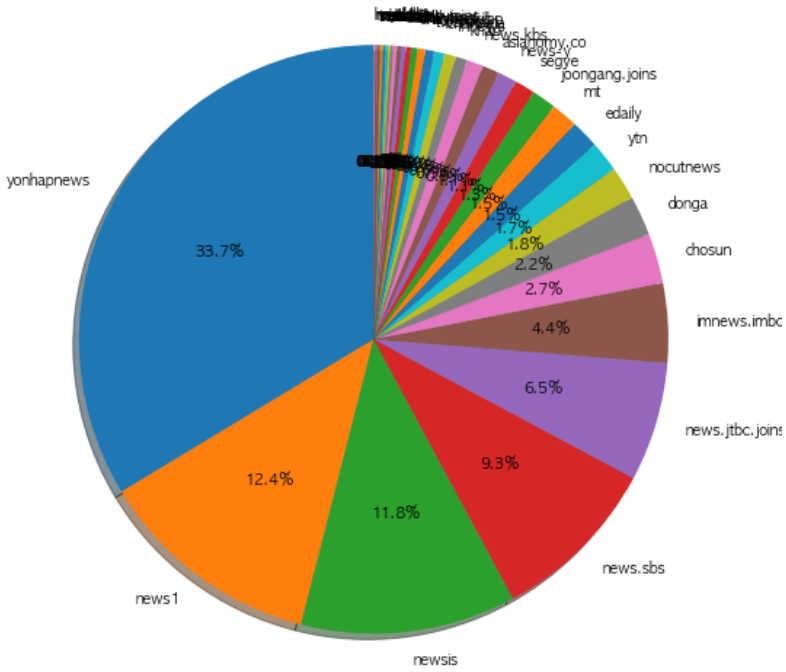


Fig. <11> The composition of press at "Naver_mHome:character-type_news_list:1:1:Y"

At the same time, it is necessary to look out the real numbers of published news and how it is possible based on the number of their journalists in the table <10>.

As there is no any needs to examine every media, the number of news and journalists implies that the policy of news layout in Naver is not natural.

Table <10> The numbers of press at "Naver_mHome:character-type_news_list:1:1:Y"

Press company name	news_id	Jounalist name
yonhapnews	479	352
news1	176	173
newsis	168	167
news.sbs	133	84
news.jtbc.joins	93	83
imnews.imbc	62	53
chosun	39	0
donga	31	0
nocutnews	26	12
ytn	24	6
edaily	22	18
mt	21	17
joongang.joins	19	19
segye	16	1

In summary, as it is mentioned at the first time in this part, sometimes the fact that we have believe or we think it is a common sense could be wrong. And this study has examine this mistakes of the common sense by using data analysis.

As a result, the tables <9, 10>, the figures <5, 6> shown us there is something unnatural on Naver's homepage. Because when we think normally without any data, it seems that there is no any problems on it. But as the several tables and figures shown us, it can be found that there is something that has to be examined.
Even though Naver is still insisting that they are just the internet portal service provider, there is no any proof why they provides 33.7% of the occupancy time on an important location just only for the single media. The answer of this question is up to the public because if the public does not have any doubt and does not request the answer to Naver, they are not going to answer this question never ever.

3.2. Cluster & Classifier

3.2.1. The general review of the classification

3.2.1.1. Distribution of predicted values

As mentioned in the section 1.5 , it is very important to obtain a predicted result with high reliability. In order to do that, it is recommendable to test enough types of algorithms, to test by using several dependant variables or to test a variety of types of data. Maybe all of them. (See the section 2.4.2)

Based on the training data from the already created clusters, classify the test(untrained) data.

The result of the classifier has a total of 263 cases as follows.

Number of cases = Number of clusters * Number of test subjects 3 (News title, News body, Tweets)

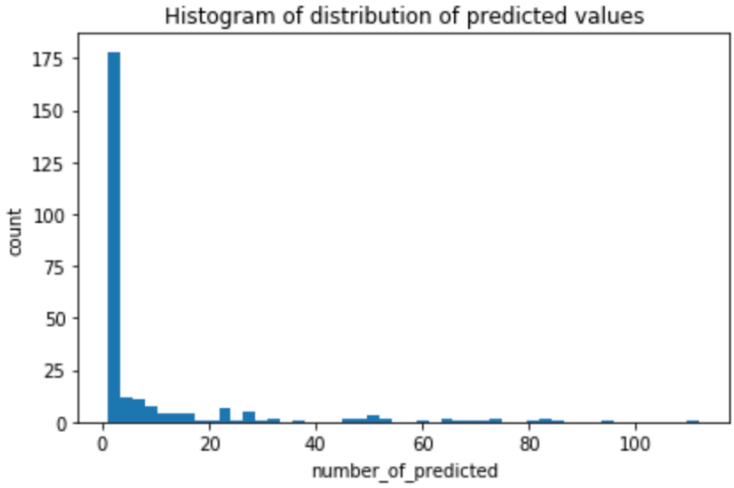


Fig. <12> Histogram of distribution of predicted values

In the above histogram, the X-axis is the number of classification prediction values, and the Y-axis is the sum of the classification results corresponding to the classification prediction value.

In the distribution chart, it can be seen that the result of the classifier's work on the test data is very unbalanced and the number of classification predictions is mostly in the range of 1 ~ 10.

For example, if a cluster is created with 9000 samplings and 1800 clusters (the number of clusters), and a classifier classifies the test objects using this cluster, It may be reasonable to classify as many as possible. There are many variables that cause the same result like as the figure <12>, depending on what clusters are used or what type of the classifier is used.

In addition, the test data used actually may be extremely biased. However, since it is used almost 140,000 news items as a test data, the researcher of this study can get a point that this possibility is remarkably low.

3.2.1.2. Highest classification results

The table <11> below is an information of clusters and classifiers from two classification results, top-ranked among the 263 classifiers' work results.

clst_algorithm	KMeans	KMeans
clustered_column	news body	news title containing only nouns
clustered_table	News	News analyzed by ETRI lang API
number_of_clusters	1000	1800
number_of_sampling	5000	9000
run_time(sec)	37.6535	122.672
clf_alorithm	MultinomialNB	MultinomialNB
test_column	news body	news body
test_table	News	News
object_count_by_predicted	[21, 9, 4, 2, 33, 1, 9, 5, 1, 13, 11, 2, 3, 22...	[1, 27, 61, 110, 32, 347, 10, 36, 7, 35, 1, 7,...
predicted	[2, 6, 24, 26, 32, 34, 37, 39, 44, 48, 49, 50,...	[0, 9, 21, 34, 37, 43, 61, 66, 68, 78, 81, 99,...
number_of_predicted	112	95

Table <11> Information of top two clusters & classifiers

having the highest classification results

The table <11> shows the most reasonable classification results when the test table / column matches the clustered table / column. This seems obvious. The news body in the news table has been sampled and clustered by the same objects using the results of the clusters so that the results should surely be the best.

Let's look at the next point of the table <11>. On the right-side result, it is shown that the clustered table & column are different. This separate analysis table is based on the language analysis using the ETRI AI API. Since the data of the language analysis is vast, it can not utilize all the information, and it is a table that extracted only nouns in consideration of the characteristics of Korean. This is because the "Term Frequency times Inverse Document Frequency" technique being used in the unsupervised-clustering technique is the core technology of the cluster. Currently, scikit-learn does not provide an analysis API for Korean language. In other words, context analysis is impossible. If so, then we can not but recognize the Korean language based on the frequency of the words. In order to overcome these limitations, it is necessary to use ETRI AI API to extract only nouns from news headlines and news texts and to make them into separate clustering target tables. Since there is a significant number of stop-words in the long text of the news text, and the API provided by scikit-learn provides stop-words corpus only for European languages, so that stop-words [22] in Korean can be recognized as plain words .

In addition to the existing scikit-learn API, the combined use of multiple APIs to overcome the characteristics of Korean and existing API constraints has also shown positive results in the study.

3.2.1.3. Comparison of the results when used ETRI API

As shown in the figure <13>, it can be seen that the distribution on the left using ETRI AI API is more uniform than that on the right. In particular, if we look at the part where the X axis value indicating the number of classification prediction results is large, it can be considered that the hypothesis that the classification prediction result value can be more reasonably obtained when clusters are extracted by only nouns is proved to some extent.

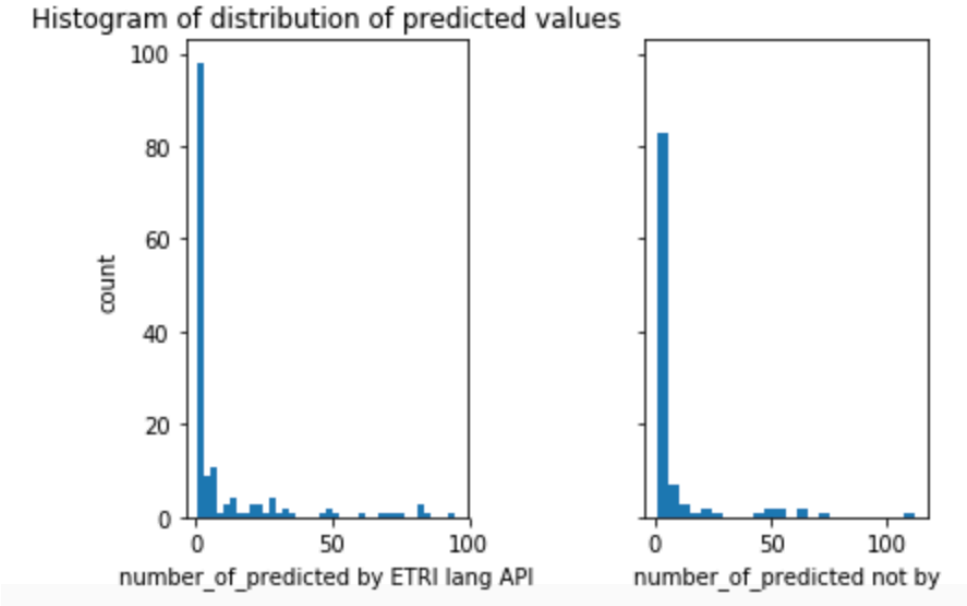


Fig. <13> The predicted distribution difference between the use and nonuse of ETRI language analysis API

3.2.2. Optimized choice of cluster & classifier

We discussed before about the reliability problem of the results by the cluster and the classifier. Many clusters had been created in a number of different cases by the two cluster algorithms and by how they are used.

In addition, it could be obtained the classification's results by a number of channels by applying two classification algorithms and its application method. The number of cases could be increased to the infinity depending on which clusters are classified using any classifiers. However, in this study it is not a good idea to use many cluster algorithms and many classification algorithms. Rather, it would seem more efficient to find a few optimized cluster-classifier combinations that classified the news and the tweets by expanding the number of cases and its usage.

3.2.2.1. The worst combinations of cluster & classifier

A total of 130 clusters and a classifier using the clusters were estimated to total 780 cases. The key idea is to find cases where one cluster produces reasonable classification predictions for all of the two classifieds (news, tweeters).

3.2.2.1.1. Effects of Homo/Heterogeneous sampling

Normally, it is known if a correlation is strong, the objects have the similarity. Not surprisingly, this simple theory can be applied to this study. The training data through the sampling and the test data being going to be applied by the clusters from the training data may be quite similar. However, in this case, more reasonable classification prediction results can be obtained when training data and test data are sampled from the different groups.

The following the figure <14> shows the result.

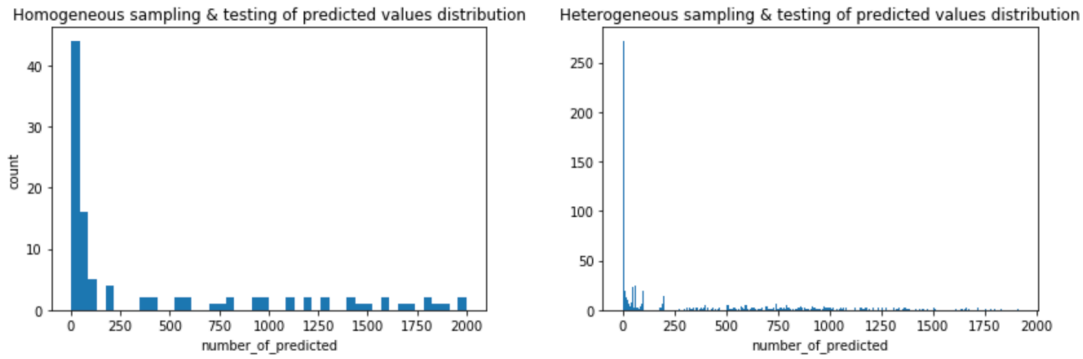


Fig. <14> Difference distributions of predictions between Homogeneous and Heterogeneous

As you can see in the figure <14>, we can see that sampling train data from the homogeneous group on the left and clustering it into the clusters derived therefrom is more reasonable than from the heterogeneous group on the right. However, both of these results were overwhelming, which is a general phenomenon occurring in

this study. This is because it was created from the various clusters when the clusters were created.

3.2.2.1.2. Sample metadata having worse predictions

The following table <12> shows 5 samples which have worse results of the classification.

	110	125	45	71	70
clst_algorithm	MiniBatchKMeans	KMeans	KMeans	MiniBatchKMeans	MiniBatchKMeans
clustered_column	news title	news title	news title	news body	news title
clustered_table	News	News	News	News	News
number_of_clusters	50	67	1000	1600	1600
number_of_sampling	1000	1000	5000	8000	8000
run_time(sec)	0.189477	0.460695	5.55164	109.205	7.49792
clf_alorithm	MultinomialNB	MultinomialNB	MultinomialNB	MultinomialNB	MultinomialNB
test_column	news title	news title	news title	news body	news title
test_table	News	News	News	News	News
object_count_by_predicted	[14011]	[14011]	[14011]	[13937]	[14011]
predicted	[5]	[7]	[5]	[207]	[111]
number_of_predicted	1	1	1	1	1

Table <12> 5 subordinates with poor cohort classification

The last number, number of predicted, indicates that there is only one classified group. If we look at the input variables that result in poor results, we can see that the MultinomialNB classification algorithm is used for the whole 5 cases. This classification algorithm results in significantly bad overall results in this study. One unusual thing is that the title of the news is often used as training data. It can be seen that the classification prediction results are not good when using data such as a news title with a short sentence length at the time of a cluster creation.

This phenomenon could be obvious given the "Term Frequency times Inverse Document Frequency" [23] concept described earlier in the cluster part. The shorter the length of the sentence, the less frequently the same word is used repeatedly.

3.2.2.2. The best combinations of cluster & classifier

Now it's time to find the most optimized combination. Let's find a cluster-classifier combination that excludes the worst possible combinations being mentioned above and classifies the news and the tweets most reasonably.

Every single day the news will not be talking about just one topic. This point is a common sense even if it is not proved. Rather, it is really strange that there is only one topic among a variety of the news on the homepage. It is reasonable to think that the classification of news topics is diverse. Therefore, it would be useful to find classification results in which the classification prediction values classified by the classifier are distributed in various ways. In this study, the contents of the news title, the body text, and the tweets had been analyzed, so that it is necessary to search for the most common classification predictions by testing each classification and by finding a few common metadata of the classification predictions among the clusters.

By excluding 105 prediction results from the worst cluster & classifier combination among 783 classification prediction results, 678 extraction targets can be obtained.

3.2.2.2.1. Top-ranked 10 combinations

The table <13> shows the results for 10 ranked among 678 classified results. The priorities are listed from left to right by sorting in reverse order as shown in the table. What is impressive is that the classification algorithm is all Perceptron. Sadly, however, Twitter being tested in the future was not ranked on the top 10 below.

clf_alorithm	Perceptron	Perceptron	Perceptron	Perceptron	Perceptron	Perceptron	Perceptron	Perceptron	Perceptron	Perceptron
clst_algorithm	KMeans	KMeans	KMeans	KMeans	KMeans	KMeans	KMeans	MiniBatchKMeans	MiniBatchKMeans	KMeans
clustered_column	news body containing only deduplicated nouns	news body	news body containing only nouns	news body containing only deduplicated nouns	news title containing only nouns	news body containing only nouns	news body containing only deduplicated nouns	news body containing only nouns	news title containing only nouns	news body containing only nouns
clustered_table	News analyzed by ETRI lang API	News	News analyzed by ETRI lang API	News analyzed by ETRI lang API	News analyzed by ETRI lang API	News analyzed by ETRI lang API	News analyzed by ETRI lang API	News analyzed by ETRI lang API	News analyzed by ETRI lang API	News analyzed by ETRI lang API
number_of_clusters	2000	2000	2000	2000	2000	1800	1800	2000	2000	2000
number_of_predicted	1911	1830	1801	1781	1755	1722	1718	1670	1661	1660
number_of_sampling	10000	10000	10000	10000	10000	9000	9000	10000	10000	10000
object_count_by_predicted	[644, 225, 216, 211, 206, 186, 141, 114, 112, ...	[182, 99, 73, 66, 64, 62, 58, 57, 52, 51, 50, ...	[2660, 643, 289, 199, 161, 139, 136, 133, 119,...	[364, 126, 120, 83, 78, 77, 70, 69, 63, 60, 59...	[3731, 1575, 96, 91, 81, 79, 63, 57, 56, 53, 4...	[738, 591, 447, 401, 127, 122, 120, 90, 80, 74...	[539, 438, 429, 167, 155, 135, 134, 134, 133, ...	[3021, 1264, 237, 209, 149, 127, 121, 107, 99,...	[1713, 1410, 253, 193, 145, 99, 98, 84, 83, 79...	[3530, 137, 82, 81, 64, 60, 59, 58, 56, 56, 55...
predicted	[463, 1195, 75, 26, 170, 1011, 1669, 1073, 170...	[636, 569, 744, 1059, 1691, 1685, 52, 258, 477...	[531, 477, 1863, 1565, 778, 1899, 1121, 372, 3...	[231, 1580, 1352, 979, 268, 242, 81, 392, 121,...	[20, 110, 1354, 465, 28, 46, 1286, 1067, 30, 1...	[1607, 1004, 655, 13, 1080, 1535, 837, 940, 91...	[1448, 1409, 391, 1636, 676, 1322, 93, 1365, 4...	[1034, 1833, 7, 1504, 1201, 72, 631, 1237, 237...	[210, 9, 592, 616, 1105, 1143, 220, 199, 214, ...	[531, 935, 22, 886, 906, 1091, 1550, 1276, 129...
test_column	news body	news title	news body	news title	news title	news body	news body	news body	news title	news title

Table <13> Optimized Cluster - Classifier Combination Top 10

Therefore, it should be re-considered to find an optimized cluster-classifier combination that includes up to the twitter.

Now as mentioned above, we are going to look at each target data.

3.2.2.2.2. News title

The table <14> shows the metadata and classification values of the classification prediction result for the news title. Here too, we can find that the various variables are similar to the above-mentioned twitter classification results.

clst_algorithm	KMeans	KMeans
clustered_column	news body	news body containing only deduplicated nouns
clustered_table	News	News analyzed by ETRI lang API
number_of_clusters	2000	2000
number_of_sampling	10000	10000
run_time(sec)	161.973	107.22
clf_alorithm	Perceptron	Perceptron
test_column	news title	news title
test_table	News	News
object_count_by_predicted	[182, 99, 73, 66, 64, 62, 58, 57, 52, 51, 50, ...	[364, 126, 120, 83, 78, 77, 70, 69, 63, 60, 59...
predicted	[636, 569, 744, 1059, 1691, 1685, 52, 258, 477...	[231, 1580, 1352, 979, 268, 242, 81, 392, 121,...
number_of_predicted	1830	1781

Table <14> Optimized combination of Cluster - Classifier for News title

3.2.2.2.3. News body

In a same way for the news body, the table <15> can be obtained.

clst_algorithm	KMeans	KMeans
clustered_column	news body containing only deduplicated nouns	news body containing only nouns
clustered_table	News analyzed by ETRI lang API	News analyzed by ETRI lang API
number_of_clusters	2000	2000
number_of_sampling	10000	10000
run_time(sec)	107.22	244.705
clf_alorithm	Perceptron	Perceptron
test_column	news body	news body
test_table	News	News
object_count_by_predicted	[644, 225, 216, 211, 206, 186, 141, 114, 112, ...	[2660, 643, 289, 199, 161, 139, 136, 133, 119,...
predicted	[463, 1195, 75, 26, 170, 1011, 1669, 1073, 170...	[531, 477, 1863, 1565, 778, 1899, 1121, 372, 3...
number_of_predicted	1911	1801

Table <15> Optimized combination of Cluster - Classifier for News body

3.2.2.2.4. Twitter

Finally, the table <16> shows the top two results of classification prediction for Twitter. In the table, most classification predictors are the same, but there is a difference between the number of training data samples and the number of classification groups for sampling.

clst_algorithm	KMeans	KMeans
clustered_column	news body containing only deduplicated nouns	news body containing only deduplicated nouns
clustered_table	News analyzed by ETRI lang API	News analyzed by ETRI lang API
number_of_clusters	2000	1800
number_of_sampling	10000	9000
clf_alorithm	Perceptron	Perceptron
test_column	text	text
test_table	twitter_UserTimeline	twitter_UserTimeline
object_count_by_predicted	[3060, 1174, 1097, 836, 514, 437, 352, 302, 26...	[3017, 1001, 926, 790, 633, 416, 367, 249, 227...
predicted	[231, 295, 170, 1083, 1669, 1722, 169, 1425, 4...	[352, 134, 1310, 1322, 938, 1030, 111, 391, 49...
number_of_predicted	1459	1382

Table <16> Optimized combination of Cluster - Classifier for Tweets

3.2.2.2.5. Final optimized combinations

Surprisingly, it was easy to find the commonality between the results of three targets. And the matching rate of the metadata of the cluster is very high. These points are checked and available in the table <17>.

clst_algorithm	KMeans	KMeans	KMeans
clustered_column	news body containing only deduplicated nouns	news body containing only deduplicated nouns	news body containing only deduplicated nouns
clustered_table	News analyzed by ETRI lang API	News analyzed by ETRI lang API	News analyzed by ETRI lang API
number_of_clusters	2000	2000	2000
number_of_sampling	10000	10000	10000
clf_alorithm	Perceptron	Perceptron	Perceptron
test_column	news body	news title	text
test_table	News	News	twitter_UserTimeline
object_count_by_predicted	[644, 225, 216, 211, 206, 186, 141, 114, 112, ...	[364, 126, 120, 83, 78, 77, 70, 69, 63, 60, 59...	[3060, 1174, 1097, 836, 514, 437, 352, 302, 26...
predicted	[463, 1195, 75, 26, 170, 1011, 1669, 1073, 170...	[231, 1580, 1352, 979, 268, 242, 81, 392, 121,...	[231, 295, 170, 1083, 1669, 1722, 169, 1425, 4...
number_of_predicted	1911	1781	1459

Table <17> Found an optimized combinations of Cluster-Classifier

The table <17> is a result table where the common points from three objects was found appropriately. All we have to do is find the common cluster-classifier combination condition of the top ranked objects in each object. Surprisingly, they all have the same metadata except for the classification predictions and the test data. It was mentioned earlier that the cluster technology is the core technology of this study. In addition, we can confirm that the advanced algorithm of the cluster algorithm proposed by the present researcher has been selected. In other words, our own algorithm, which complements the characteristics of Korean and the limitations of existing algorithms, yields more efficient and reasonable results.

3.3. News

3.3.1. Composition of News by Press

3.3.1.1. All press

First, the figure <15> shows how the news posted on Naver homepage and main page of news service are distributed by media companies during a certain period.

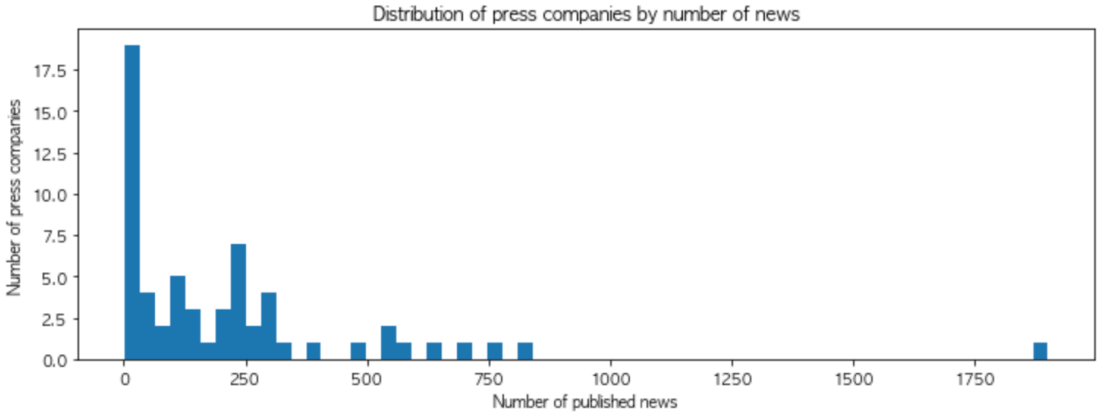


Fig. <15> Distribution of press companies by number of news

This is the histogram of the number of journalists according to the number of news published. As you can see at a glance, one company has uniquely posted a large number of news. And the second-ranked media company posted half of the news. As you can see from this fact, Naver, an internet portal company, can not say that it has fair editorial rights of news as a press, but it is fair itself.

Let's look at the composition ratio more specifically in the following figure <16>.

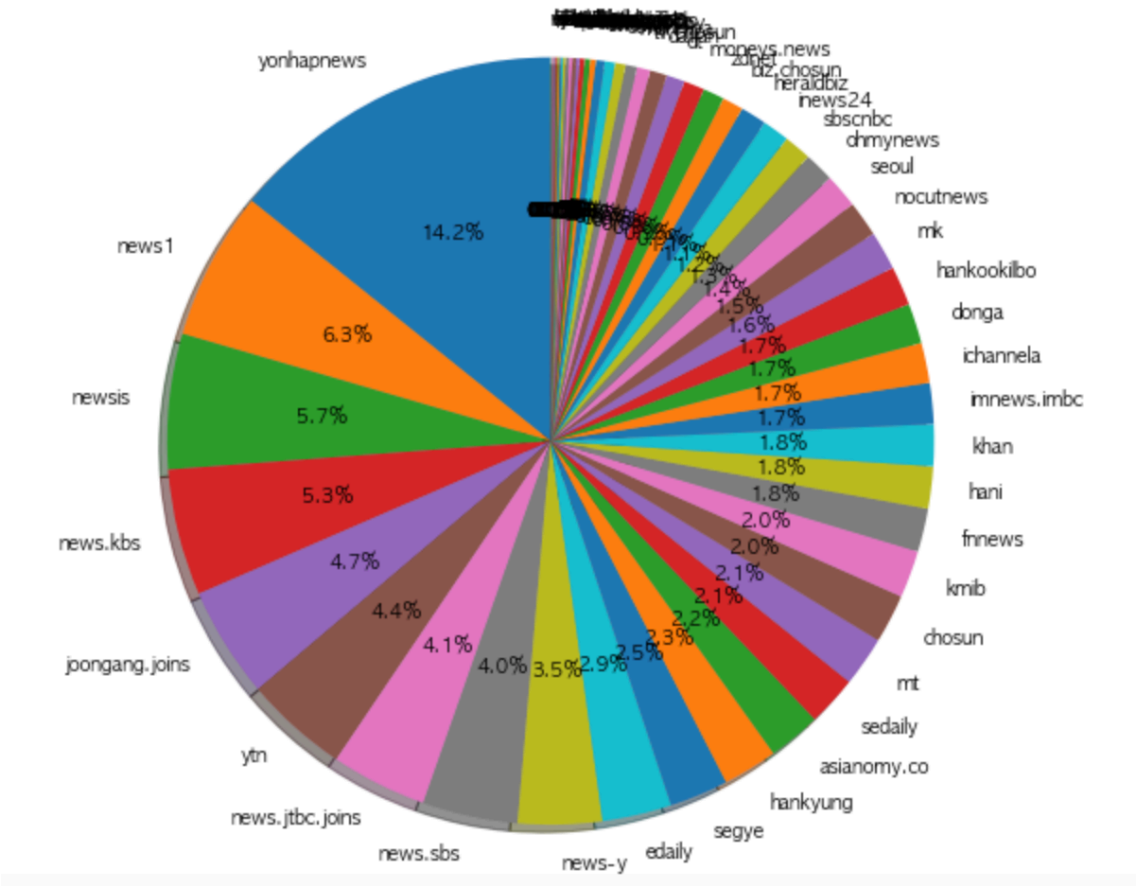


Fig. <16> Composition of the press about whole published news

Among the 67 news outlets, only 61 have posted news on Naver's core web pages, of which the *yonhapnews* accounted for a staggering 14.2 percent. It is more than twice as big as the amount of the news by *news1(The second raking in the pie graph)*.

3.3.1.2. Top 10 press

The table <18> on the right-side shows only the top 10 press' list with the absolute number of the published news. This table <18> is different from the previous table <10> in aspects of that the table <10> just treated some news which were in the specific location "Naver_mHome:character-type_news_list:1:1:Y".

press_company_name	news_id
yonhapnews	1901
news1	836
newsis	761
news.kbs	712
joongang.joins	624
ytn	584
news.jtbc.joins	541
news.sbs	536
news-y	473
edaily	391

Table <18> The number of news of top 10 about whole published news

From now on, it should be wondered how many news are published per day on average. In other words, the issue is how many news just one press can release a day. Because it is natural for all journalists to cover articles and to write articles when a social issue arises, but experience has proven that the publication of the article was sometimes unreasonable.

The figure <17> shows the number of news releases over time in the above-mentioned top 10 media companies over a period of time. It can be seen from the graph that the *yonhapnews* press attracts a lot of attention.

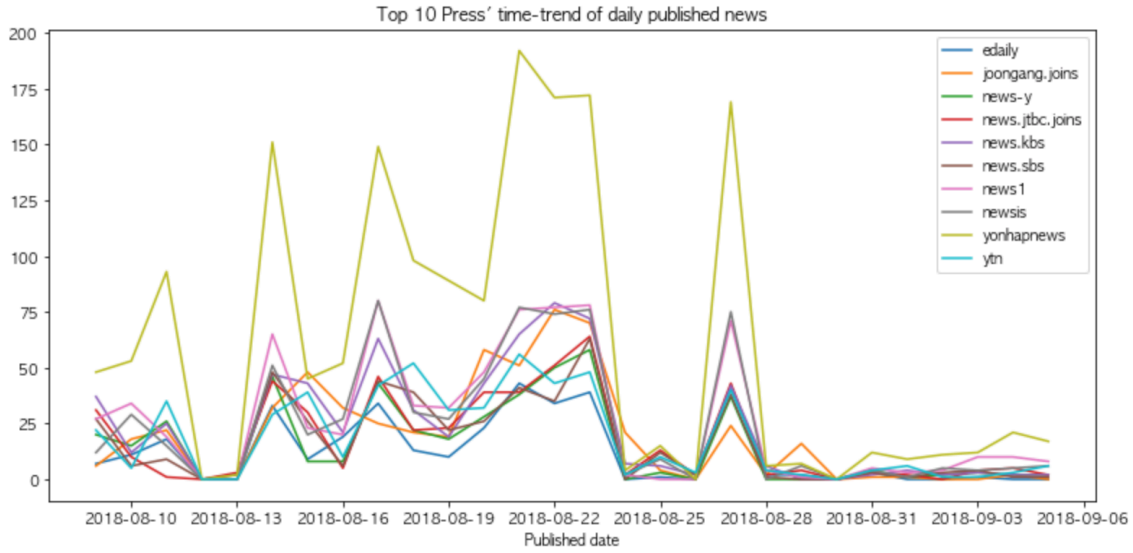


Fig. <17> Top 10 Press' time-trend of daily published news

At this point, it is necessary to look around some other statistical calculation on it. Look at the table <19>.

press_company_name	count	mean	std	min	25%	50%	75%	max
yonhapnews	28.0	59.928571	64.594994	0.0	8.50	33.0	94.25	192.0
news1	28.0	26.071429	28.771036	0.0	2.75	14.5	37.50	80.0
newsis	28.0	24.285714	28.235592	0.0	3.50	10.5	33.75	80.0
news.kbs	28.0	22.678571	25.060007	0.0	2.00	9.5	42.25	79.0
joongang.joins	28.0	19.607143	22.606684	0.0	0.75	17.0	26.75	76.0
news.jtbc.joins	28.0	17.500000	19.559027	0.0	2.00	5.0	33.00	64.0
news.sbs	28.0	16.428571	18.689329	0.0	1.00	6.0	29.00	63.0
news-y	28.0	15.500000	18.240167	0.0	0.75	6.0	26.50	58.0
ytn	28.0	18.785714	19.137625	0.0	2.75	8.0	36.00	56.0
edaily	28.0	12.285714	14.983589	0.0	0.00	5.0	20.00	43.0

Table <19> Top 10 Press' time-trend of daily published news

This press seems to be a massive media company that can produce up to 192 news articles only in a single day.

While comparing 1st to the others from 2nd to 10th in the table <19>, *yonhapnews* overwhelms other press. It is doubtful whether one media company can produce so many news in just one day.

There are just only three public press in Korea. The reason why a public media company is needed in a country is because the fairness of media reports is absolutely necessary. Public media companies operate on the taxes of the public so that they are free from advertising revenue from external companies that affect the company's durability. Contrary to this, non-public media means that the company's profits come from advertising, which makes it easier to publish the news with impaired fairness. The research shows that the vast majority of the news on the main page of Naver's news service, which is watched by the majority of the people every day, is produced by the media, which is likely to be damaged by the fairness of the media coverage.

3.3.2. Composition of News by Journalists

The question then arises whether the press, the News Union, is able to generate a large amount of news and to let them posted on Naver with the large quantities.

If this press, *yonhapnews*, is a mess media group overwhelming the 3 major public press, the doubt will be vanished. In order to figure it out, the number of journalists working in yonhapnews should be also considered.

We had already looked at the composition ratio of news press in the previous part. If it is executed to compare the two composition ratios between the press and journalists in each press company, the difference and the doubt can be easily understood.

The figure <18> is a pie chart showing the composition ratio of the total journalists by the whole press.

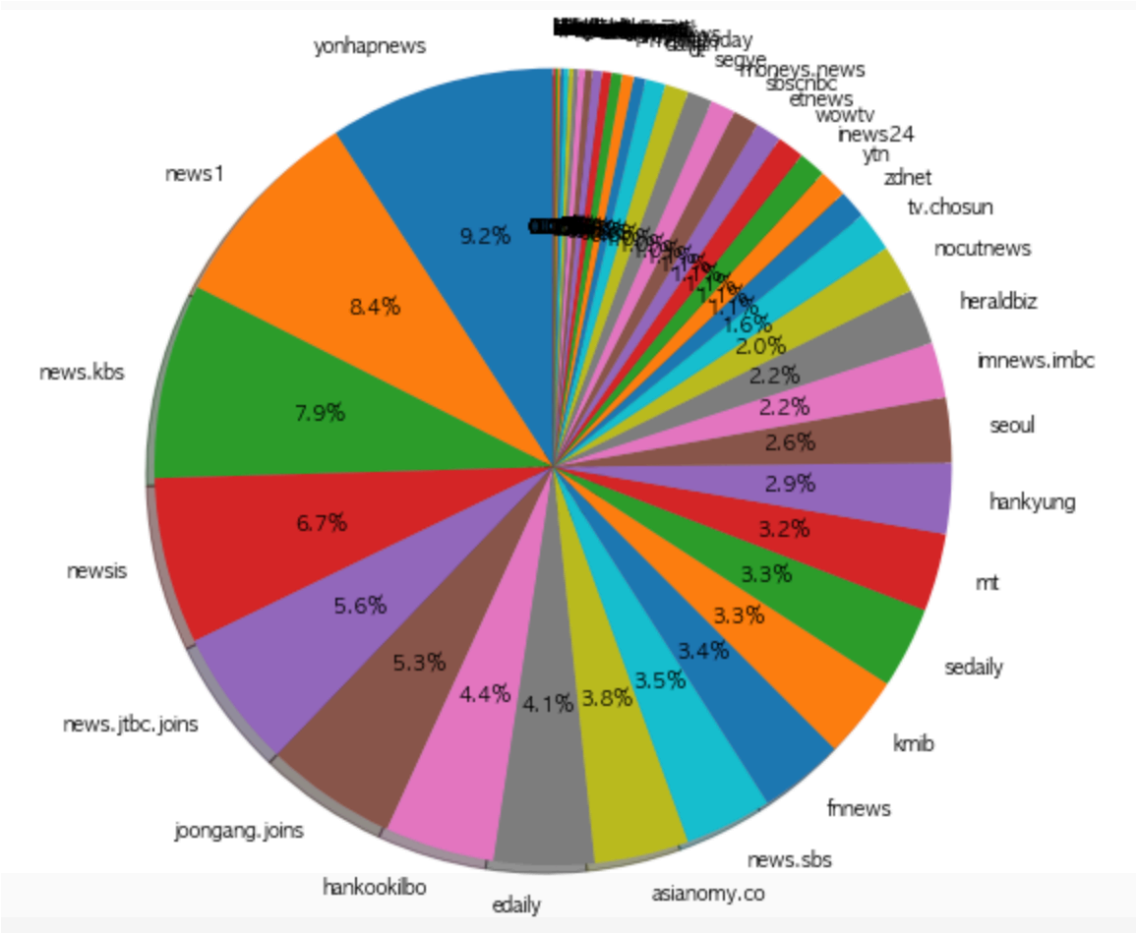


Fig. <18> Composition of the journalists each press company

Among the whole 2090 journalists, only 9.2% of journalists are from the *yonhapnews*. Nonetheless, their news volume is up to 14.2% comparing to the figure <14>.

It is need to look at the absolute number in order to compare the ratio number with, which shows the exact number of journalists in each press who wrote any articles or news and published them on Naver.

Press company name	Jounalist name
yonhapnews	192
news1	175
news.kbs	165
newsis	141
news.jtbc.joins	118
joongang.joins	111
hankookilbo	93
edaily	85
asianomy.co	80
news.sbs	74
fnnews	72
kmib	70
sedaily	69
mt	66
hankyung	60
seoul	55
imnews.imbc	47

Table <20> Number of the journalists each press company (right-side)

3.3.2.1. Top 20 journalists

In this step, we are going to dig into the top 20 journalists who published a lot. To make it clear, it is recommended to look at the figure <19> and table <21> together.

First, the figure <19> indicates the trend of daily news publication time of the top 20 journalists who issued the most news for a certain period of time(Almost one month).

Noteworthy is that *yonhapnews's* reporter Kim Seung-wook and Lee Seung-hyung wrote up to 24 articles and 20 articles in a single day. This can be more clearly seen in the table <21>.

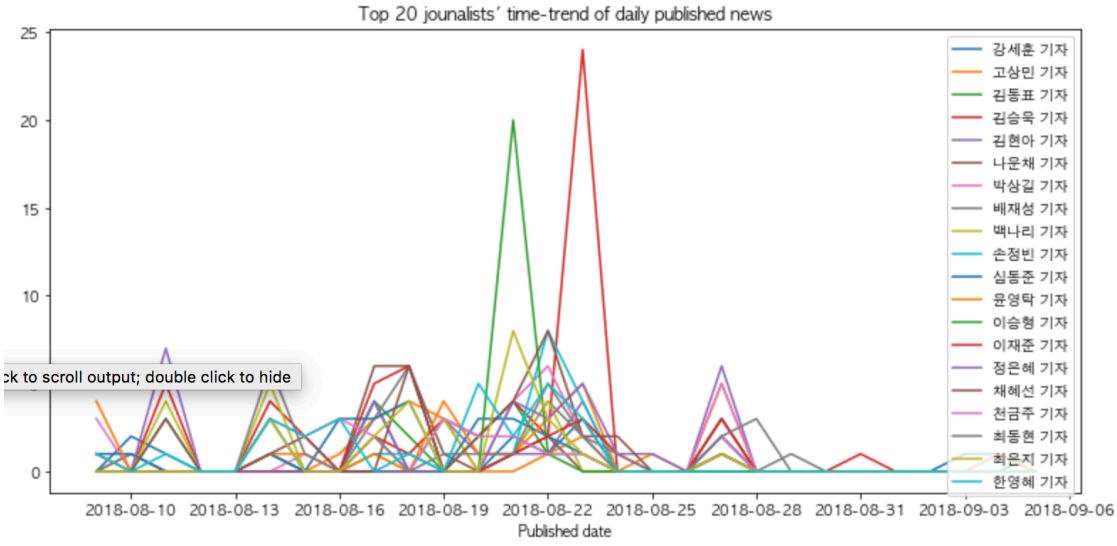


Fig. <19> Top 20 jounalists' time-trend of daily published news

언론인명	count	mean	std	min	25%	50%	75%	max
김승욱 기자	28.0	1.857143	4.576221	0.0	0.0	0.0	2.00	24.0
이승형 기자	28.0	0.785714	3.774742	0.0	0.0	0.0	0.00	20.0
최은지 기자	28.0	0.785714	1.792105	0.0	0.0	0.0	0.25	8.0
손정빈 기자	28.0	0.642857	1.725930	0.0	0.0	0.0	0.00	8.0
채혜선 기자	28.0	1.107143	2.006273	0.0	0.0	0.0	2.00	8.0
정은혜 기자	28.0	0.785714	1.771318	0.0	0.0	0.0	0.25	7.0
최동현 기자	28.0	0.892857	1.547741	0.0	0.0	0.0	1.00	6.0
김현아 기자	28.0	0.821429	1.611385	0.0	0.0	0.0	1.00	6.0
나운채 기자	28.0	0.714286	1.652319	0.0	0.0	0.0	0.25	6.0
박상길 기자	28.0	0.750000	1.669442	0.0	0.0	0.0	0.00	6.0
배재성 기자	28.0	0.750000	1.456149	0.0	0.0	0.0	1.00	6.0
이재준 기자	28.0	0.750000	1.601504	0.0	0.0	0.0	0.25	6.0
한영혜 기자	28.0	0.928571	1.537881	0.0	0.0	0.0	1.25	5.0

	count	mean	std	min	25%	50%	75%	max
count	20.0	20.000000	20.000000	20.0	20.0	20.0	20.000000	20.000000
mean	28.0	0.837500	1.785311	0.0	0.0	0.0	0.875000	7.300000
std	0.0	0.262311	0.858902	0.0	0.0	0.0	0.651415	5.252568
min	28.0	0.607143	1.065947	0.0	0.0	0.0	0.000000	3.000000
25%	28.0	0.750000	1.417299	0.0	0.0	0.0	0.250000	5.000000
50%	28.0	0.750000	1.574623	0.0	0.0	0.0	1.000000	6.000000
75%	28.0	0.830357	1.737277	0.0	0.0	0.0	1.062500	7.250000
max	28.0	1.857143	4.576221	0.0	0.0	0.0	2.000000	24.000000

Table <21> Number of News Publications Top 20 Most Daily News Publishers

Second, the average number of top 20 journalists writing the most articles is only 7.3, as it can be seen in the table <21> on the right side. Two reporters from *yonhapnews* wrote over three times the articles in one day.

In general, the journalists of the news media, *yonhapnews*, report three times more news than other press' journalists. This fact implies two possibilities. First, it could mean that the journalists of *yonhapnews* are smarter and more engaged than other journalists. Second, they would just write articles without considering fact-based researches.

3.4. Correlation between news and tweets

So far, we have been analyzing various aspects of the layout of news, the news and tweets. Finally, it is time to figure out whether there is any correlations between them. First of all, we are going to look for it about the whole events. Second of all, about the specific events.

3.4.1. About whole events

3.4.1.1. Time correlation of news and tweets

Below the Fig. <20> and Fig. <21> are the time trends in total daily numbers of news and tweets. First, figure <20> on the top applies the same scale to the two objects and expresses the mean value and the standard deviation value together to help understand the meanings in the graph. On the other hand, figure <21> at the bottom shows once again with two different scales about the same data, because the number of news is significantly higher than the number of politicians' tweets.

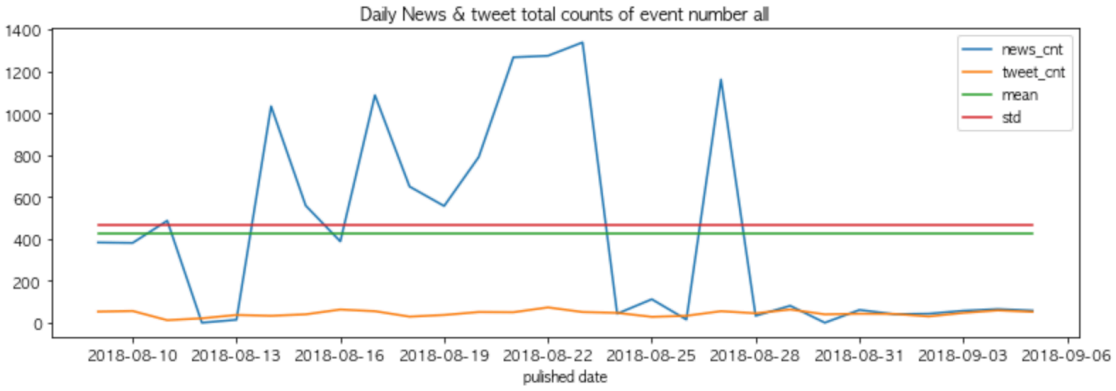


Fig. <20> Daily News & tweet total counts of all event number

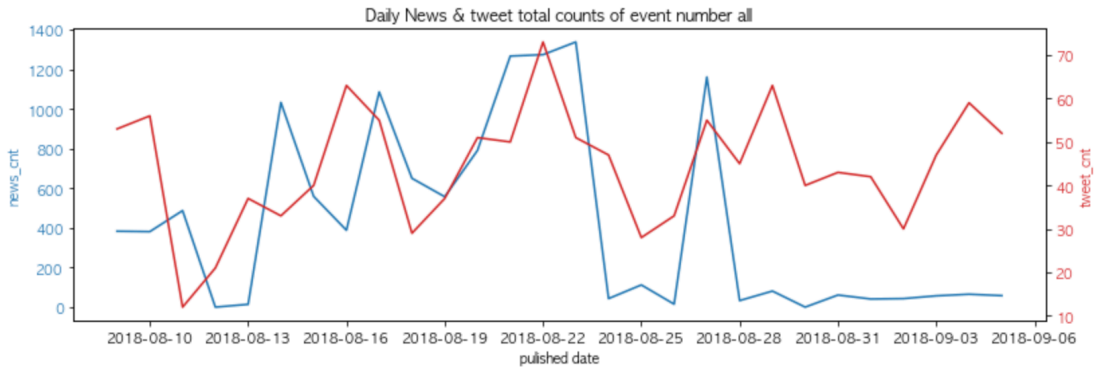


Fig. <21> Daily News & tweet total counts of all event number in different 2 scales

The news in the figures <20,21> shows the number of events that are collected at regular interval (10 ~ 30 mins) every day for a certain period. As you can see in the figures, the number of the published news had been changed rapidly by date. This can be confirmed by the fact that the standard deviation is higher than the average

value. It does not follow the naturally occurred standard distribution the number of the published news on the homepage and on the news service page of Naver.

The invisible trend line of tweets in figure <20> at the top is clearly visible in the bottom figure <21> so that it can clearly seen the relationship between the two objects.

Some suspicions can be raised while watching the figure <21>. It implies that the politicians had done a lot of tweets when the news events were published massively on Naver. Perhaps the number of news publications with large deviations from day to day may be related to this suspicion.

Once again, this data shown above in the figures <20,21> is the results from the same cluster. In other words, it means we should keep in mind that the news and the tweets share many of the same events. Considering these basic conditions, the graph at the bottom gives us a lot of insights. It is because it implies that the media and the politicians who produce the news and the tweets may be connected.

It may be natural for journalists to publish a news about what happened when it happened, and also for politicians to comment on the event and the news. But there is something we have missed. That is if they are the politicians, they have to express their opinions carefully because politicians, especially parliamentarians, are elected by the public. Nonetheless, as the graph tells, korean politicians tweet ahead as soon as the news are published.

At this moments, it seems quite reasonable to think of if there could be some connections between the korean parliamentarians and the press. But it is not still clear that it is a fact there is the obvious connections between them. Although it is sure that is not sure, but it also is sure there is something in them.

Either way, it is not good for the public. However, it is the worst thing for the public not to know about something because it could make the democracy destruction happen. Therefore, it is valuable to analyze this more deeply.

3.4.1.2. Distribution of classified predictions of news and tweets

In the previous section, we looked at the total number of news and the trends in the total number of tweets of politicians. Let's look at the distribution of the groups classified by the classifier from now on.

As can be seen from the histogram, it can be seen that the distribution of classification predictions for the two analysis subjects (news, twitter) is mostly overlapped. Of course, except for some classification values that deviate from the mean and its standard deviation in all classification predictions. This is because this phenomenon appears throughout the classification predictions by all cluster-classifier combinations.

Nevertheless, when we calculate the correlation between two objects, we can see that the hypothesis of this researcher is reasonable.

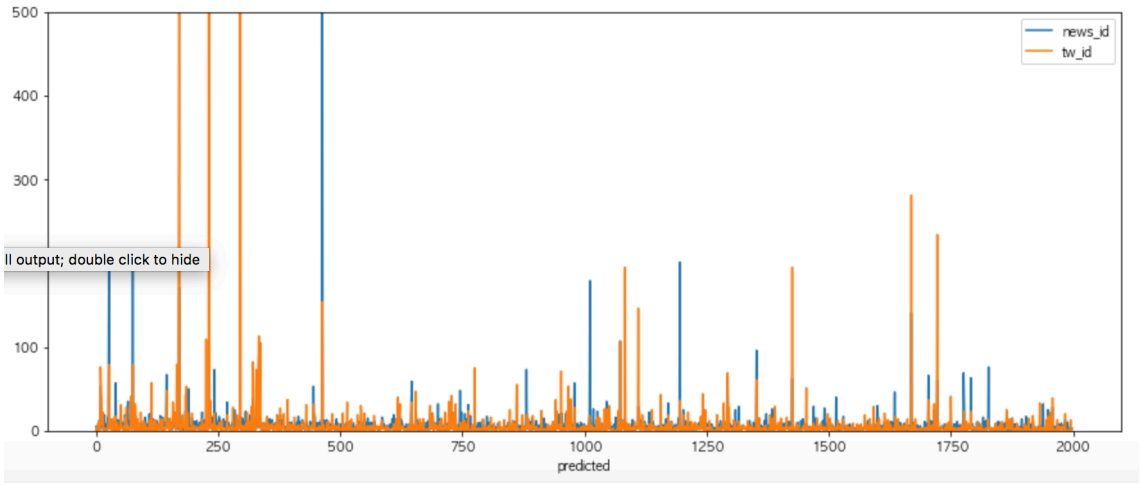


Fig. <22> Histogram of classified prediction values of news and twitter

The table <22> is about the statistical calculation result for two objects. The standard deviation (std) and the correlation coefficient from the left table are on the right-side table. The correlation value is higher than 0.21. This was calculated by Pandas' default method, which is Pearson's standard correlation coefficient.

It could seem to be low. But remember. The datas being treated in this study are not numerical datas but are the textual datas. We should consider that those numbers in the tables were derived from the textual datas by using the clustering and the classification.

Moreover, this study is based on unsupervised machine learning. The positive correlation is consistent with the purpose of this study.

If it comes out as a negative correlation, it would have been a proof to overturn the hypothesis. As regarding of this points, the number 0.21 is not low. It is quite meaningful.

	news_id	tw_id
count	1911.000000	1911.000000
mean	6.990058	7.571952
std	17.624929	52.663348
min	1.000000	0.000000
25%	2.000000	0.000000
50%	4.000000	2.000000
75%	7.000000	5.000000
max	535.000000	1823.000000

	news_id	tw_id
news_id	1.000000	0.214074
tw_id	0.214074	1.000000

Table <22> Analysis correlations between news and tweets

3.4.2. About specific events

It was mentioned that the news and the politicians' tweets were all categorized by the same clusters and classifiers, but the news was divided into a larger number of groups. In order to analyze the correlation between the two, the type and time of the event must be matched because it is mandatory to compare the journalists 'news publication and politicians' tweets about the same incident at a similar time.

Due to these limitations, unlike previous data, only the news and tweet data from '2018-08-09' to '2018-09-05' will be covered. The number of news groups classified by the original classifier is 2000, but when limited to a certain period, it is reduced to 1,193 groups. Likewise, the twitter data is originally 1,459, but it is reduced to 437. The number of news items is reduced from 13,358 to 11,920, and the number of tweets is reduced from 14,613 to 1,193.

This suggests a great deal. This is especially true in the number of groups, because not only is the number of groups of news and tweets categorized reduced, but also common events have to be extracted. As a result, the number of common events is regressively reduced to 423. This is less than 437, the number of groups in the tweet. It is a small number of events in the study, but fortunately, this paper does not cover all 423 events.

3.4.2.1. Choice of specific events

From now on, it will be explained how to extract high-correlated events for 423 events, 11,920 news stories, and 1,193 tweets.

The table <23> shows the data of each news and tweet grouped by the event number(predicted column) and then sorted in reverse order of news' counts and tweet's counts. As you can see in the table, there is no common occurrence between the two. So let's find out about 423 cases.

	news_id
predicted	
55	107
755	62
658	56
1722	52
1570	49
382	43
569	42
165	38
734	33
163	31

	tw_id
predicted	
231	141
295	100
170	33
1722	23
1669	23
169	19
328	18
1425	18
1932	18
333	14

Table <23> Number of top 10 events of News / Tweet for mutual analysis

The table <24> shows the top 10 by sorting in reverse the correlation coefficients for the time between event news and tweets. Correlation coefficients range from -1 to +1. The closer to +1, the higher the correlation in this study because the negative correlation coefficient does not have any meaning in this case, which needs a positive relation. You will remember that you did not find a common occurrence between news and tweets in the "" data / news tweets tweets in the top 10 events "". As a result of the actual calculation, we could find that there were high correlation coefficients between the two.

	corr_coeff	event_num
322	1.000000	1469
50	0.970725	187
288	0.866025	1323
61	0.782467	231
312	0.730297	1433
123	0.681385	497
185	0.609994	755
82	0.516398	321
93	0.426401	370
377	0.323070	1722

Table <24> Correlation coefficient by event between news and tweets

Then next, the figures <23,24> are the distribution of these correlation coefficients.

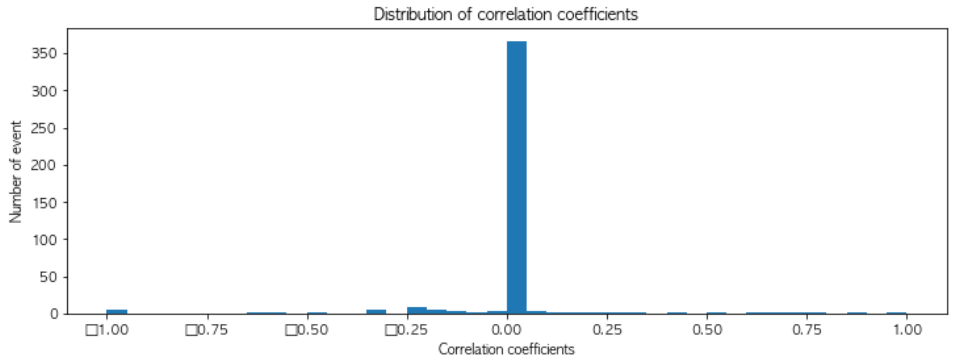


Fig. <23> Distribution of correlation coefficients

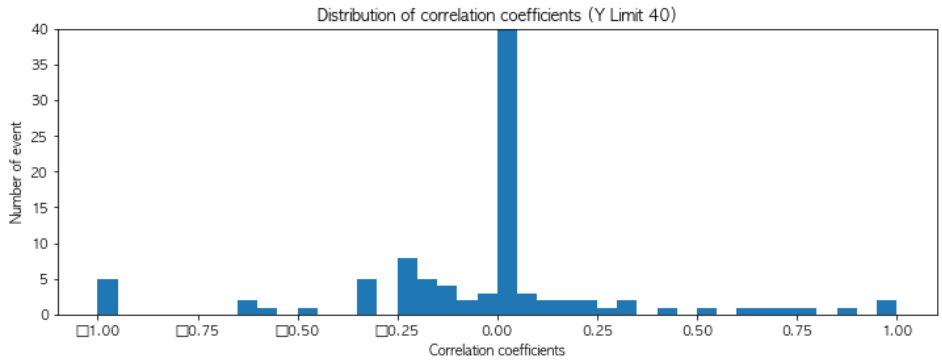


Fig. <24> Distribution of correlation coefficients (Y Limit 40, bins 20)

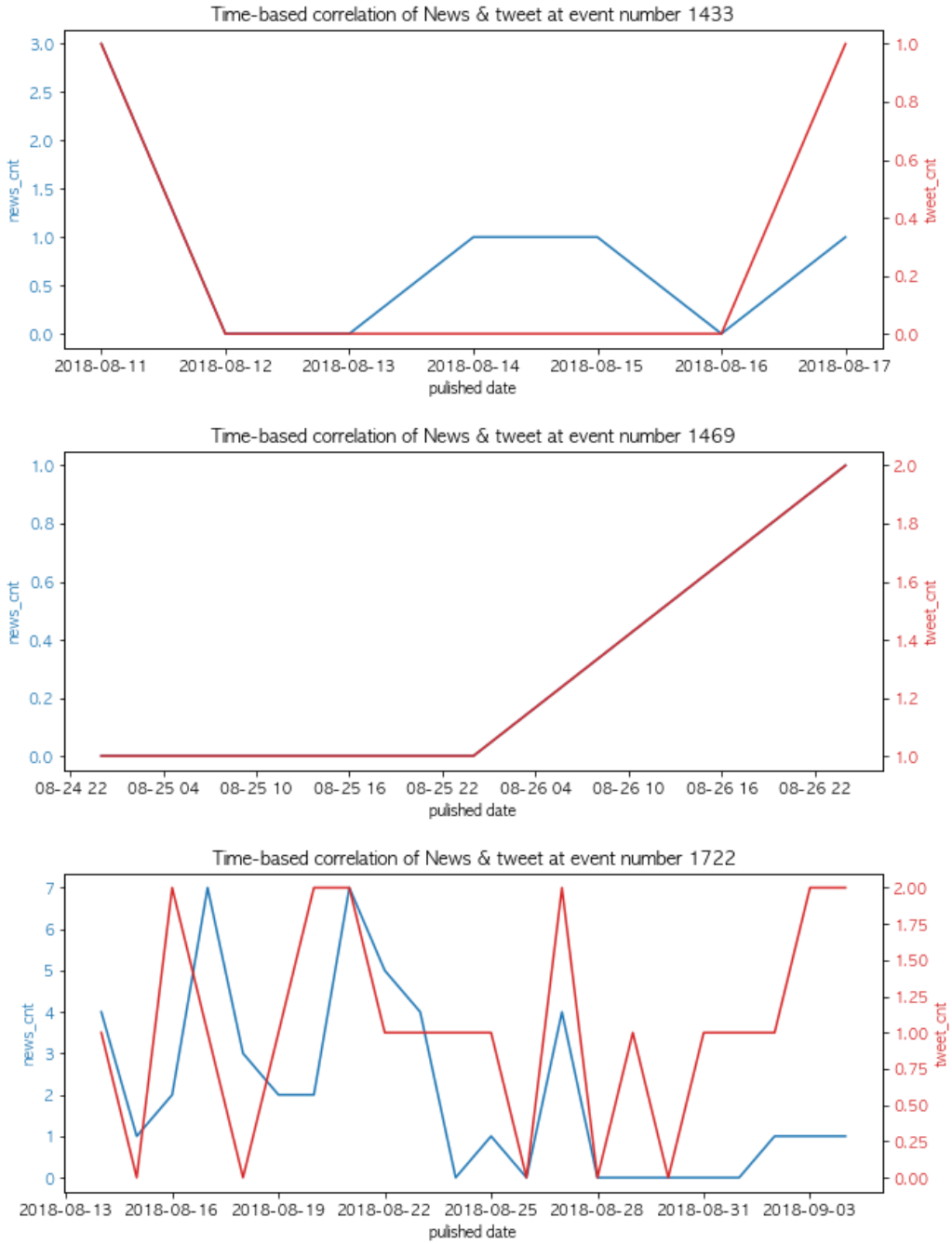
The figure <23> shows the whole, on the other hand, the figure <24> shows enlarged view of 40 Y-axis. As shown in the figure, the correlation coefficient exists in the + area as a whole. This means that journalists 'news releases and politicians' tweets have a positive correlation to the same event and time.

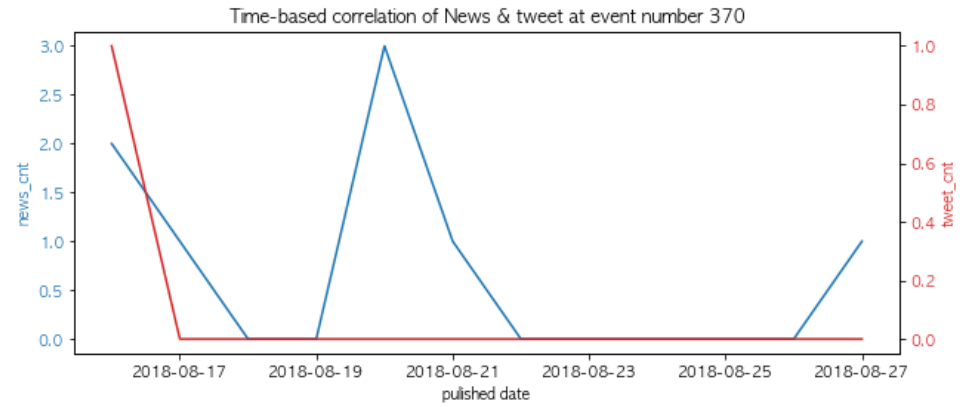
This distribution is very important in the selection of events to be analyzed at the next part. This is because it provides a basis for easily selecting events that can prove the validity of the hypotheses proposed by the researcher.

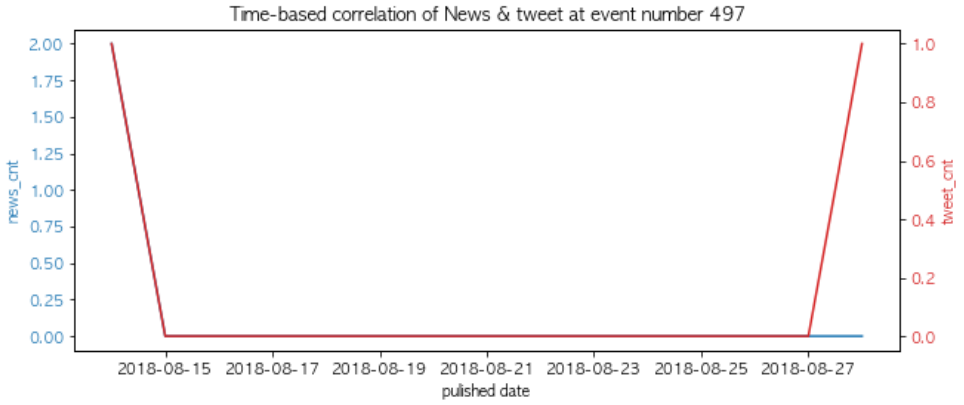
3.4.2.2. Time correlation of news and tweets

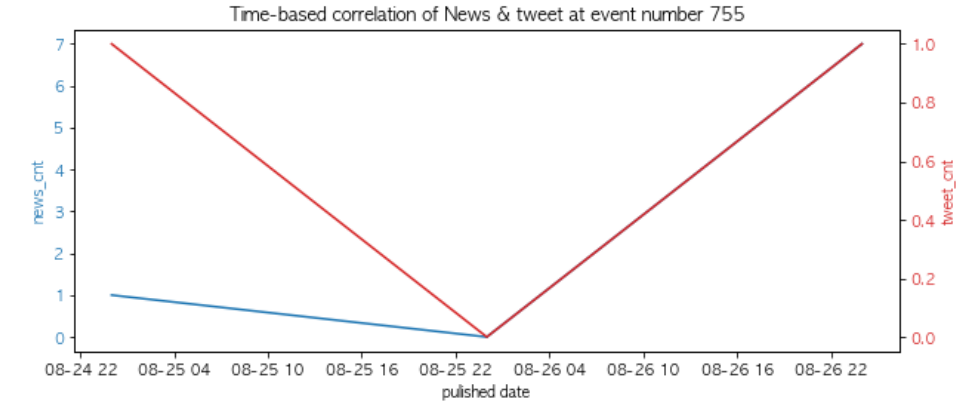
This study focuses on some events having a high correlation coefficient to highlight the relationship between the journalists' news and the politicians' tweets. Therefore, we will look at the 10 events deeply that were chosen from the table <27> "Correlation coefficient by event between news and tweets" .

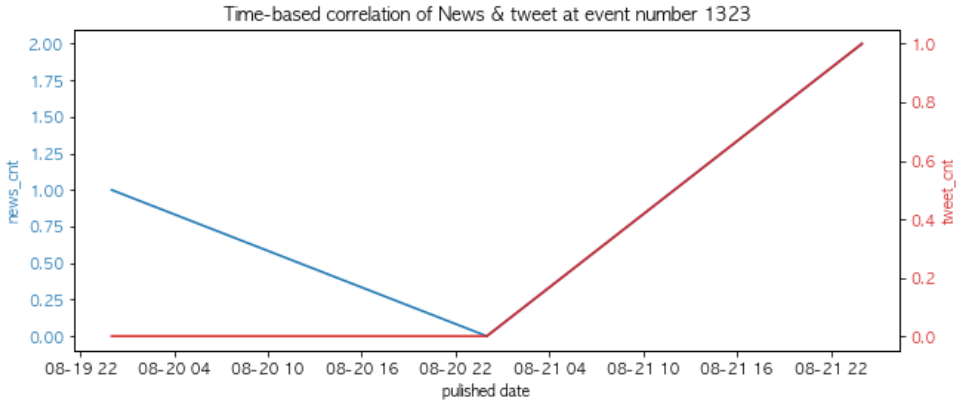
First, let's look at the daily time trends between the number of the news and the number of politician's tweets about the 10 sample events.

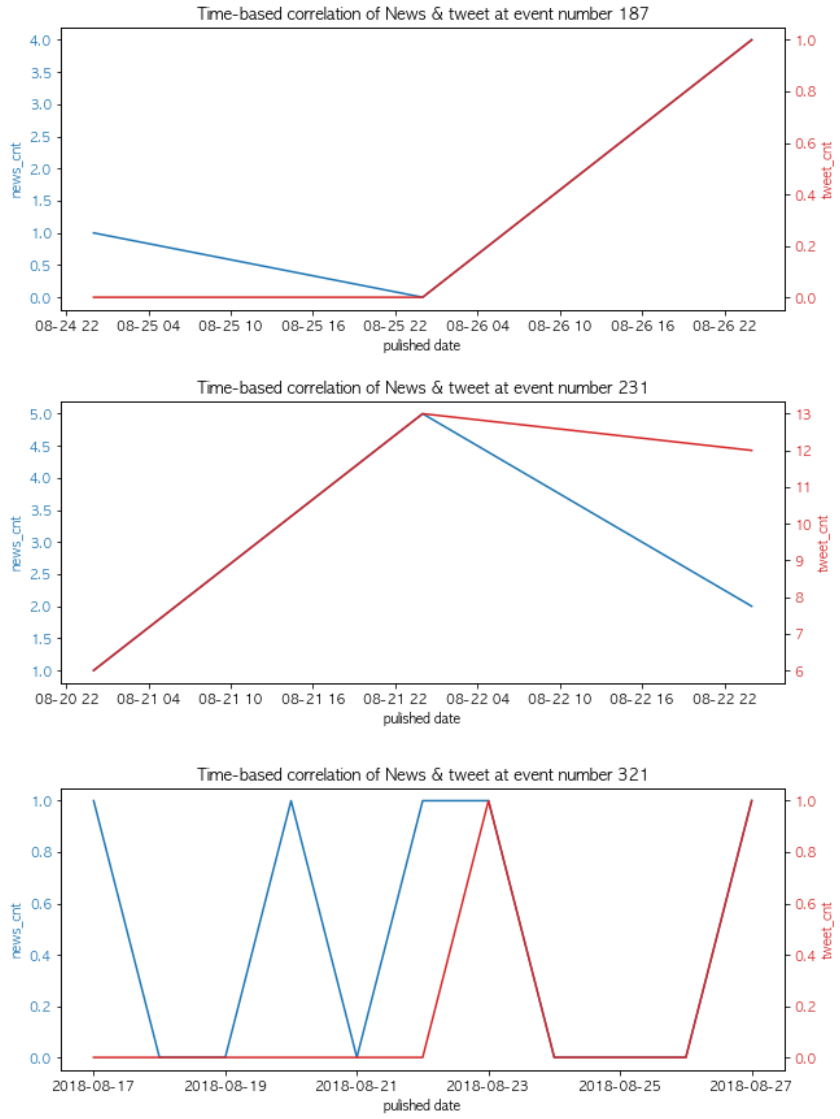



Time-based correlation of News & tweet at event number 370
news_cnt
tweet_cnt
pulished date
2018-08-17
2018-08-19
2018-08-21
2018-08-23
2018-08-25
2018-08-27


Time-based correlation of News & tweet at event number 497
news_cnt
tweet_cnt
pulished date
2018-08-15
2018-08-17
2018-08-19
2018-08-21
2018-08-23
2018-08-25
2018-08-27


Time-based correlation of News & tweet at event number 755
news_cnt
tweet_cnt
pulished date
08-24 22
08-25 04
08-25 10
08-25 16
08-25 22
08-26 04
08-26 10
08-26 16
08-26 22


Time-based correlation of News & tweet at event number 1323
news_cnt
tweet_cnt
pulished date
08-19 22
08-20 04
08-20 10
08-20 16
08-20 22
08-21 04
08-21 10
08-21 16
08-21 22



Figs. <25> Time-based correlation of News & tweet at event number N's

As you can look at the 10 graphs above, the coincidence between them has been clearer in the specific events than in the whole events. They seems like moving together as if they arrange the timing before. In particular, *event number 231*, in which there are relatively amount of tweets, shows as if two players are working together systematically. The number of politicians' tweets also tends to increase as the number of news stories increases sharply for any event number N. Although it was observed only graphs with some unusual points in a total of 413 datas, it is pretty reliable to believe that these 10 graphs are sufficient to show there is a high correlation between the news and the tweets.

3.4.3. Adherence between the press and politicians

As a final view point, this time we are going to analyze the correlations between the press and the politicians. For the first time, it is necessary to choose press and politicians having quite high correlation coefficient based on the previous results.

The following figure <26> shows the composition ratio of the media companies having a positive correlation coefficient. *yonhapnews* and *news1* got ranked #1 and #2 once again.

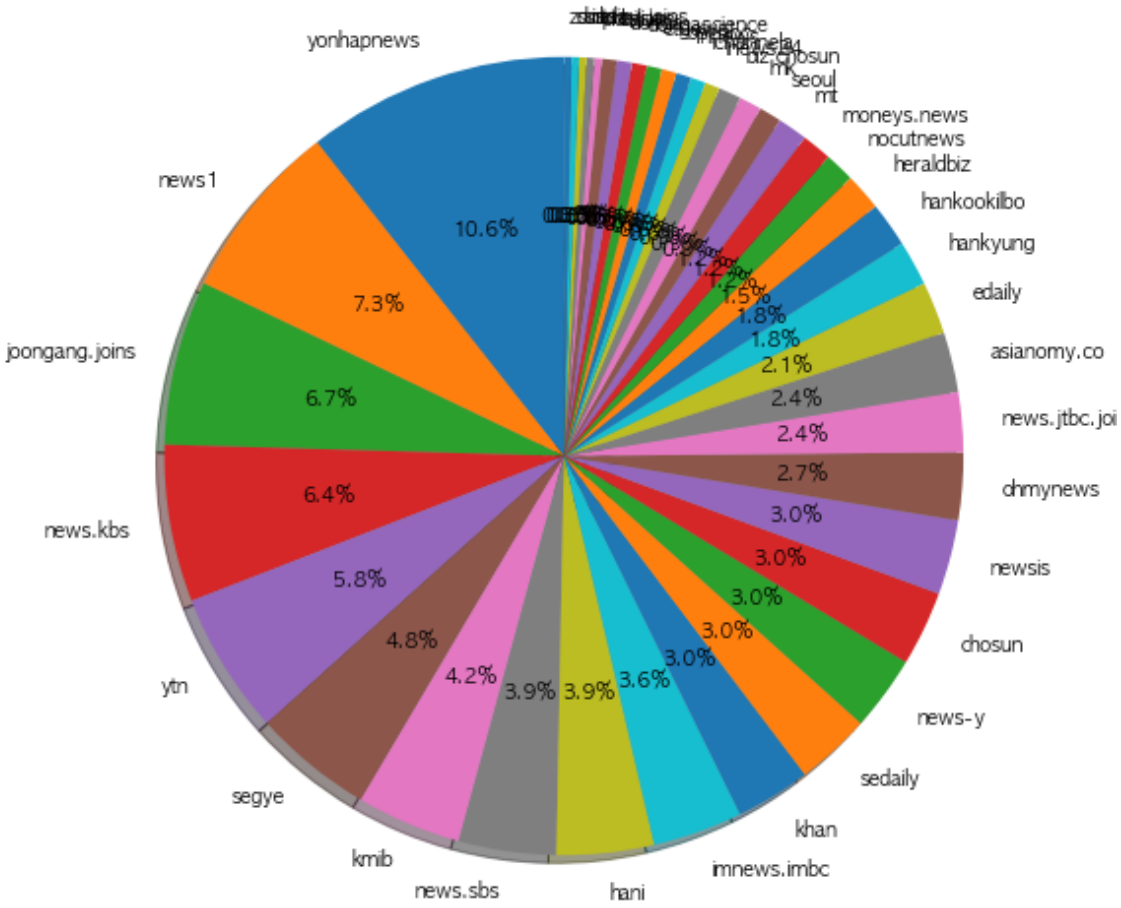


Fig. <26> Composition ratio of press companies having a positive correlation coefficient

And then next, let's look at politicians in an extension of the same way.

In the figure <27>, 37 politicians of 59 who are activated on the tweets activity had the positive correlation coefficient. (It was explained about the total assemblymen in korea in the chapter 2)

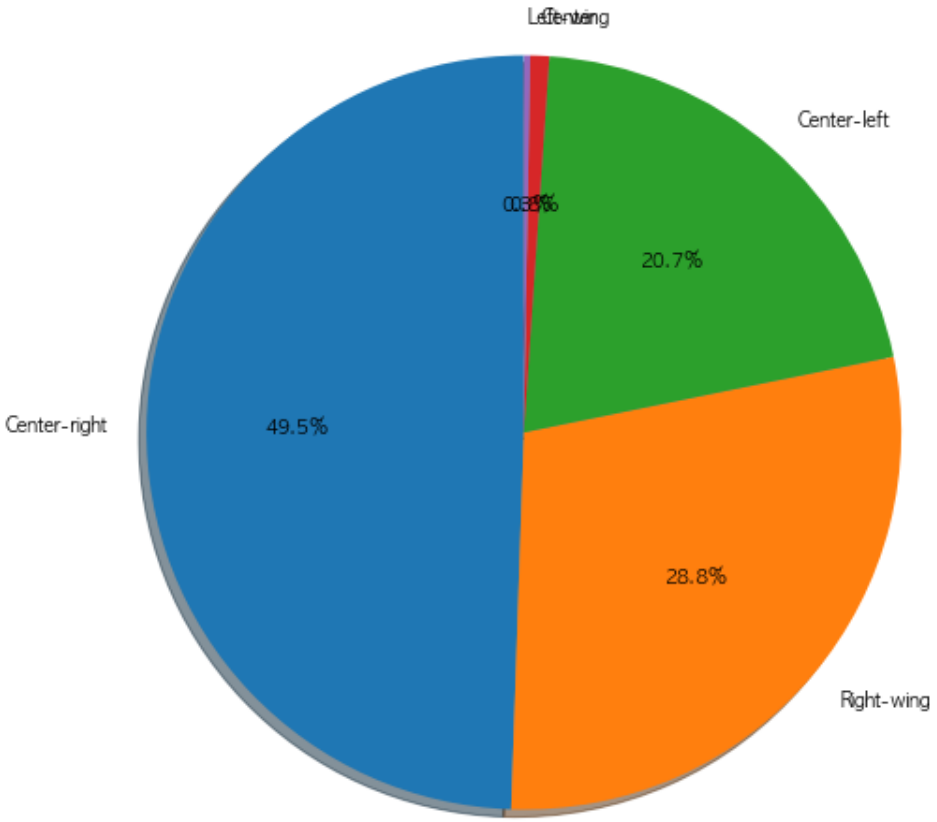


Fig. <27> Composition ratio of political spectrum having a positive correlation coefficient

As a result, most of the members of parliament who using twitter and who belong to right-wing's tendancy were coincidentally talking about the same news at the same time.

Those pie charts give us some credibility to recent results. Generally speaking, there was a kind of conspiracy theory that the conservative party lawmakers manipulate the public opinion by back transactions with the media. But there was not any concrete evidence in the past. However, as the technology of data analysis has been developing, it is possible to find out some evidences or some clues that there is a great connection between them.

4. Discussion

4.1. Difficulty of data collection

It had to use as many method as possible for collecting data and parsing data. Because each web service company has a different Response Timeout, or limits the number of requests per unit time, each variable must be calculated and set in the batch program of the system. For example, in the case of a Naver portal service, if a page is collected in 2 seconds, it is recognized as a DDOS attack and prevents service access itself from a specific IP for a while. In order to continuously monitor the research activities with these hacking factors and to improve existing vulnerabilities, internal reporting should be provided.

4.2. Software engineering on data analysis

The generic agile development methodology is useful because it allows you to get new insights that you did not expect in the initial design of the entire data analysis process, which requires new source data or a series of complex and regressive processes This is because it repeats countless times. The researcher himself has repeated the process of collecting, analyzing, and reporting three steps repeatedly through this project, and has gone through the process of erasing some existing source codes. The water-fall way, which is the old-fashioned development methodology and which is still used in the software industry field in Korea at the current, is the methodology that is absolutely incompatible with data analysis projects.

5. Conclusion

5.1. Secondary goal 1

It is to find out how Naver arranges the news on their homepage and on their news service pages. Through out the subchapter 3.1 , it could be found that Naver let a few news stay in the important locations longer. The differences of the occupancy times do not seem to be random, natural or spontaneous. This detection does not tell us for sure that Naver is manipulating the public opinions by controlling the occupancy time of each news in a location. However, it is obvious that there is something suspicious on them after analyzing the Naver's news arrangement policy.

5.2. Secondary goal 2

The third goal is to find out what kinds of correlations are there between the press and the politicians. This came from just private insights by lots of experiences in real life. There is still not any evidences, research reports in korea. That was exactly why I wanted to figure it out by myself. Through out the subchapter 3.3~4, the insights by experience had a more reliable credibility even though it will have to be improved more. The more important lesson is that this study taught us if there is something unclear it is needed to be analyzed by real data not by his own thought or feeling.

5.3. Secondary goal 3

The second goal is about how a bunch of datas related to the news can be analyzed programmatically with a high reliability. In fact, it is not possible to manually analyze amount of the news datas every day. In the period of collecting data, which was almost a month, maximum 1338 news were published in a particular day. Hence, it is inevitable to use Machine learning to classify them not manually. Through the subchapter 3.2, we could find that adopting some technologies of machine learning is quite available to analyze the textual data like news even though its reliability is still in a process of developing.

Moreover, we could have a better results of using machine learning by mixing several different libraries. This was inevitable because scikit-learn library is especially on a purpose of European langueges. That is why this study had to apply some libraries and APIs about Korean language. In addition to that, the researcher's insight, which is just to use the nouns, made the results of the clustering better while considering the korean language's characteristics.

This goal is like a platform for performing other goals and in consequence t h e results from machine learning technology was quite enough to derive a meaningful results of the entire study.

5.4. Main goal

The main purpose of this study is to find some social issues about the media environment in korea. As a one of the citizens in korea, this purpose could have made me more motivated. We can not conclude that Naver manipulates the public opinion, we could nevertheless say there is something wrong on it. To recognize this view point is more important than to justify who is guilty or not. Because the better our society will be the more people are there recognizing what is wrong.

As a result, this study resulted in a quite good outcome when considering that this is just a first step of a long journey.

6. Bibliography

[1] 오승훈 박태우. (2018). 네이버, 모든 뉴스 손대며 언론 아니다? 전문가들 “뉴스 편집서 손 떼야”. The Hankyoreh. http://www.hani.co.kr/arti/economy/it/843815.html. Accessed 12 December 2018.

[2] Marta Fernández Diego. (2017). Chapter 1. Introduction to Business Analytics. MUGI, BAN. https://drive.google.com/open?id=14nXw69m9nsdCeJSHdmIlij03Eh5QfzuU. Accessed 28 September 2017.

[3] Naver’s developer center . https://developers.naver.com/main/. Accessed 2 December 2017.

[4] Pandas. http://pandas.pydata.org/pandas-docs/stable/generated/pandas.read_html.html. Accessed 8 March 2017.

[5] Requests: HTTP for Humans. http://docs.python-requests.org/en/master/. Accessed 8 May 2017.

[6] Twitter Open API. https://developer.twitter.com/content/developer-twitter/en.html. Accessed 10 May 2018.

[7] Tweepy. http://www.tweepy.org/. Accessed 10 May 2018.

[8] Beatuiful Soup. https://www.crummy.com/software/BeautifulSoup/. Accessed 8 May 2017.

[9] NLTK. https://www.nltk.org/. Accessed 8 April 2018.

[10] KoNLPy. http://konlpy.org/en/latest/. Accessed 30 April 2018.

[11] Eunjeong L. Park, Sungzoon Cho. (2014). KoNLPy: Korean natural language processing in Python. 제26회 한글 및 한국어 정보처리 학술대회 논문집. http://dmlab.snu.ac.kr/~lucypark/docs/2014-10-10-hclt.pdf. Accessed 29 March 2018.

[12] Ministry of Science and Technology. https://www.msit.go.kr/. Accessed 2 January 2018.

[13] ETRI (Electronics and Telecommunications Research Institute). https://www.etri.re.kr/eng/main/main.etri. Accessed 2 January 2018.

[14] Public artificial intelligence open API · DATA service portal. http://aiopen.etri.re.kr/. Accessed 2 January 2018.

[15] Scikit-learn. http://scikit-learn.org/stable/index.html. Accessed 13 December 2017.

[16] Pandas. https://pandas.pydata.org/. Accessed 1 December 2017.

[17] Matplotlib. https://matplotlib.org/. Accessed 5 December 2017.

[18] Jupyter Notebook. http://jupyter.org/. Accessed 29 September 2016.

[19] scikit-learn developers. (2017-2018). Scikit-learn Machine Learning Map. http://scikit-learn.org/stable/tutorial/machine_learning_map/index.html. Accessed 15 June 2018.

[20] scikit-learn developers. (2017-2018). K-Means equation. https://scikit-learn.org/stable/modules/clustering.html#k-means. Accessed 16 June 2018.

[21] Wikipedia. (2018). Perceptron equation. https://en.wikipedia.org/wiki/Perceptron. Accessed 14 June 2018.

[22] Wikipedia. (2018). Stop words. https://en.wikipedia.org/wiki/Stop_words. Accessed 3 January 2018.

[23] Wikipedia. (2018). Term Frequency times Inverse Document. https://en.wikipedia.org/wiki/Tf–idf. Accessed 3 January 2018.

[24] 권도경, 손기은. (2018.06.11). 네이버 ID 0.7%가 하루 30만개 댓글 달며 여론 주도. 문화일보. http://www.munhwa.com/news/view.html?no=2018061101031303325001. Accessed 12 December 2018.

[26] National Assembly. 대한민국국회. http://www.assembly.go.kr/assm/userMain/main.do. Accessed 10 March 2018.

[26] 노경진. (2018.05.10). 네이버 '뉴스 편집 중단'...댓글 기능 언론사가 결정. MBCNEWS. https://youtu.be/Ec_VOEFWIEA. Accessed 12 December 2018.

[27] Stephen Evans. (2016). Will South Korean president Park Geun-hye be impeached?. BBC News. https://www.bbc.com/news/world-asia-38170132. Accessed 12 December 2018.

[28] Jonathan Marcus. (2018). Trump Kim summit: What did it actually achieve?. BBC News. https://www.bbc.com/news/world-us-canada-44484322. Accessed 12 December 2018.

[29] 한국일보. (2017.10.22). 뉴스배치 조작 드러난 네이버 언론으로서 공공성 자각해야. 한국일보. http://www.hankookilbo.com/News/Read/201710221938162790. Accessed 12 December 2018.

[30] May Lee. (2017). Debunking the Korean Search Engine Market Share in 2017. What We Can (After All This) Reasonably Say About the Market Share. The Egg. http://www.theegg.com/seo/korea/korean-search-engine-market-share-update-2017/. Accessed 12 December 2018.

[31] 방송통신진흥본부 방송통신기획부. (2014). 인터넷 뉴스 이용자 조사 (pp83-83). 한국방송통신전파진흥원. https://www.kca.kr/open_content/bbs.do?act=file&bcd=research&msg_no=171&file_no=1. Accessed 12 December 2018.

[32] 김인철. (2017). “종이신문 위기"...정기구독률 1996년 69.3%→2016년 14.3%. 한국경제신문. http://news.hankyung.com/article/2017040421198. Accessed 12 December 2018.