SPECIFIC KEYWORD BASED
NEWS ARTICLES MULTIPLE CLASSIFICATION

Kyungwon Kim*1,2 and Kyoungro Yoon2
1 Korea Electronics Technology Institute, Republic of Korea
(E-mail: kwkim@keti.re.kr)
2 Konkuk University, Republic of Korea
(E-mail: kwstone@konkuk.ac.kr, yoonk@konkuk.ac.kr)

ABSTRACT – The industry analysis service evaluates the industry using both structured and unstructured data for
the industry. Utilization of unstructured data is increasingly used in evaluation and trend analysis for specific industries.
In order to use news articles which are representative unstructured data for trend analysis, it is necessary to classify
the news articles into corresponding industries. In this paper, we propose a method of multiple classification of news
articles into industry by using industry-specific keywords.

Keywords: Industry evaluation, Industry-specific keyword, News article, Multiple classification

1. INTRODUCTION
The existing industry analysis service evaluates the relevant industries based on the formal data such as
financial, non-financial, risk, and credit of a company corresponding to the target industry group. In addition,
we utilize unstructured data analysis such as news articles for the industry trends and future prospects. In
order to use news articles which are representative unstructured data for industrial analysis, it is necessary
to classify the collected news articles into the corresponding industries.
In this paper, we propose a method for multiple classification of news articles collected in a specific industry
based on industry-specific keywords for industry analysis.

2. INDUSTRY-SPECIFIC MULTIPLE CLASSIFICATION OF NEWS ARTICLES
To classify news articles into multiple categories by industry, a document classifier is needed. A typical
document classifier classifies documents mechanically by applying predefined classification criteria. In
recent years, machine learning techniques have been used to learn classification models using seed datasets
and classify real data by applying learned models. Classifier based on machine learning requires good seed
data for classification learning. However, it is impossible to obtain tagged seed data by multiple
classification of news articles by industry. We generate industrial classification criteria based on the
industry-specific keywords generated by industrial experts and compare the news articles with the main
keywords included in the collected news articles.

2.1. Industry-specific keywords pre-processing
Korea National Statistical Office publishes about 2,000 Korean Standard Industrial Classification(KSIC)
based on International Standard Industrial Classification(ISIC) of United Nations Statistics
Division(UNSD)[1]. We classify collected news articles with the industry-specific keywords based on
KSIC. The keywords of industrial classification are composed of compound nouns. We classify them into
atomic words and assign different weights to each keyword. Through this process, the weights of the
keywords according to their importance are determined for each industry.

2.2. Weighted keyword based new article multiple classifier
We calculate industry weighting for the main keywords included in the collected news articles, based on
the preprocessed industry keyword dictionary. Key words included in a news article are weighted according
to their importance in the article. Keyword-specific weightings of news articles are estimated based on the
TF-IDF algorithm[2]. The formula for calculating the industry classification weight of the corresponding
news article based on the keyword weight of the news article and industry keyword weight is as follows.
The news article $d_i$ to be analyzed is composed of the keyword $k_{j,d_i}$, and the industry classification keyword $k_j$ includes the classification weight $w_{k_j,c}$ for the industry C. Therefore, the industrial classification weight of the news article $d_i$ is obtained by multiplying the appearance frequency $CNT_{k_j,d_i}$ of the keyword $k_{j,d_i}$ included in the news article $d_i$ by the industry weight $w_{k_j,c}$ of the keyword $k_j$, and then summing up by industry. If the industry classification weight for each news article is derived through the above process, it can be refined according to the utilization purpose. It can be filtered news articles that have a small industry impact by taking into the industry classification weight distribution of news articles. In addition, it can be used for analysis by using only the top ranking of the industry classification of the news article.

In this paper, we exclude news articles that industry classification weights are below the threshold in order to filter articles that are considered to have low influence on the industry. The industry classification weighting threshold is determined by reflecting the distribution of the industrial classification weights of all articles to be analyzed. We applied the SOFTMAX function to only the top five industry classification weights by news articles to make the filtered news articles comparable.

### Table 1. Result of the news article multiple classification

<table>
<thead>
<tr>
<th>Article id</th>
<th>List of industry classification weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>88</td>
<td>C29100:0.2486/C10700:0.2486/R91100:0.1892/C33900:0.1892/J58100:0.1244</td>
</tr>
<tr>
<td>105</td>
<td>S94900:0.2290/P85600:0.2214/O84200:0.2214/M71500:0.1679/N75900:0.1603</td>
</tr>
<tr>
<td>125</td>
<td>O84100:0.3878/O84200:0.2041/R90100:0.1429/R91100:0.1429/I56100:0.1223</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>5605766</td>
<td>O84100:0.4800/M71100:0.1760/O84400:0.1280/O84200:0.1120/S94900:0.1040</td>
</tr>
</tbody>
</table>

### 3. CONCLUSION AND FUTURE WORKS

In this paper, we propose an industrial classification method for classified news articles based on industry classification keywords. We refined the keyword weights for each industry through preprocessing, and applied the keyword weights in the collected news articles and industrial classification keyword weights. We filtered out news articles with low impact for industry classification based on the industry weight distribution by news articles. Finally, we applied the SOFTMAX function to only the top five industry classification weights by news articles. Using the industry multiple classification method of news articles proposed in this paper, it will be possible to analyze the industry using unstructured data such as news articles.

### ACKNOWLEDGMENT

This work was supported by the Technology Innovation Program (20000195, Development on the BigData Analysis platform technology for IoT Home appliance product of small and medium sized home appliance company) funded By the Ministry of Trade, Industry & Energy(MOTIE, Korea)

### REFERENCES

1. UN, “Handbook of Input-Output Table Compilation and Analysis”, 1999.