Journal of the Korean Physical Society, Vol. 76, No. 5, March 2020, pp. 368∼377

Multi-Label Classification of Historical Documents by Using Hierarchical Attention Networks

Dong-Kyum Kim* and Byunghwee Lee*
Department of Physics, Korea Advanced Institute of Science and Technology, Daejeon 34141, Korea

Daniel Kim
Merck Sharp and Dohme Korea, Seoul 04637, Korea

Hawoong Jeong†
Department of Physics, Korea Advanced Institute of Science and Technology, Daejeon 34141, Korea
Institute for the BioCentury, Korea Advanced Institute of Science and Technology, Daejeon 34141, Korea and
Asia Pacific Center for Theoretical Physics, Pohang 37673, Korea

(Received 3 September 2019; accepted 17 September 2019)

The quantitative analysis of digitized historical documents has begun in earnest in recent years. Text classification is of particular importance for quantitative historical analysis because it helps to search literature efficiently and to determine the important subjects of a particular age. While numerous historians have joined together to classify large-scale historical documents, consistent classification among individual researchers has not been achieved. In this study, we present a classification method for large-scale historical data that uses a recently developed supervised learning algorithm called the Hierarchical Attention Network (HAN). By applying various classification methods to the *Annals of the Joseon Dynasty* (AJD), we show that HAN is more accurate than conventional techniques with word-frequency-based features. HAN provides the extent that a particular sentence or word contributes to the classification process through a quantitative value called 'attention'. We extract the representative keywords from various categories by using the attention mechanism and show the evolution of the keywords over the 472-year span of the AJD. Our results reveal that largely two groups of event categories are found in the AJD. In one group, the representative keywords of the categories were stable over long periods while the keywords in the other group varied rapidly, exhibiting repeatedly changing characteristics of the categories. Observing such macroscopic changes of representative words may provide insight into how a particular topic changes over a historical period.

PACS numbers: 01.75.+m, 07.05.Mh, 07.05.Kf
Keywords: Deep learning, Recurrent neural network, Text analysis, Big data, History
DOI: 10.3938/jkps.76.368

I. INTRODUCTION

Since the information technology revolution, various cultural heritages that used to be accessible only to skilled professionals have been computerized, and now anyone can browse the related literature through the World Wide Web. As computerized large-scale literature datasets are released, various analytical methodologies have been developed to analyze them quantitatively and systematically. In particular, state-of-the-art machine learning (ML) approaches have influenced numerous academic disciplines such as politics, social science, and history. Such eclectic ML applications have provided new opportunities for integrating both qualitative and quantitative research methodologies [1,2].

As the sheer size of some historical texts exceeds the reading capability of individual researchers, individuals have difficulty carrying out macroscopic analyses that cover the entire range of a given text. Therefore, interpretations among researchers are inevitably different, even for the same historical events, resulting in both advantages and disadvantages for qualitative research. From this point of view, text analysis based on ML techniques offers a consistent and reproducible approach that is essential for systematic studies on large-scale texts.

For instance, Google has digitized more than 15 million books (or 12% of all books published since 1450) by using

*These authors equally contributed to this work.
†E-mail: hjeong@kaist.edu

optical character recognition methods; then, it investigated the frequencies of consecutive words. In one study, through an N-gram analysis of the period 1800–2000, Google quantified such features as changes in grammar over time, the reputation of jobs, memory effects, and censorship [3]. In a similar work, Klingenstein *et al.* analyzed the semantic clustering of words used in the court records of the Old Bailey, the British Central Court, from 1760 to 1913 [4] and identified changes in the perceptions of violence over time by measuring changes in the lexicon used to describe violent and nonviolent crimes. Such analyses using large-scale bibliographic data have contributed to the understanding of macro- and time-based cultural phenomena.

Recently, beyond analysis based on simple word usage, a variety of ML- and artificial intelligence-based text analysis models have been actively developed. In particular, deep learning models using recurrent neural networks (RNNs) featuring long short-term memory [5] have shown remarkable performance in various natural language processing tasks, such as machine translation [6], sentiment analysis [7], and text classification [8,9].

Classifying historical texts into different topics and genres helps to search literature efficiently to identify important event subjects of a particular age. As mentioned, large-scale historical documents have, to date, been classified by various historians because the vast amount of information the texts contain makes them impossible to be handled by one person. This leads to inconsistent classification results for articles with the same content. Figure 1(a) demonstrates this, showing how articles with the same content can be classified differently by year. This problem of inconsistency is an obstacle to quantitative historical analysis.

Such text classification is one of the main applications where ML algorithms can be applied to historical text analysis. Conventional text classification models have word-based features, but in recent years, RNN-based deep learning models that take into account the whole context achieve the most advanced performance. The strength of deep learning lies in its ability to learn functions directly from data that map the input text to the related output in an end-to-end fashion, without requiring all prior knowledge to be input into the model. Among these models, the Hierarchical Attention Network (HAN) [10] classifies documents by modeling the underlying hierarchical structure of a text. In addition, a quantitative value in HAN, *attention*, shows how much a particular sentence or word contributes to the classification.

In the meantime, most quantitative research has been conducted on western literature, with relatively fewer studies on the historical texts of eastern countries [11–13]. In this work, we quantitatively analyze the categorical characteristics of historical events by using a large-scale Korean historical document: the *Annals of the Joseon dynasty* (AJD) [14]. The AJD, registered as a UNESCO Memory of the World, is one of the most representative historical records in Korea, covering a period of 472 years (1392–1863). The AJD is suitable for viewing socio-cultural history because it was carefully recorded over a long period of time in a regular manner with exact timestamps, a feature rarely found in similarly sized western texts. The National Institute of Korean History has classified the articles of the AJD into 40 categories that indicate the event topics of each article (see Section II, *Data*).

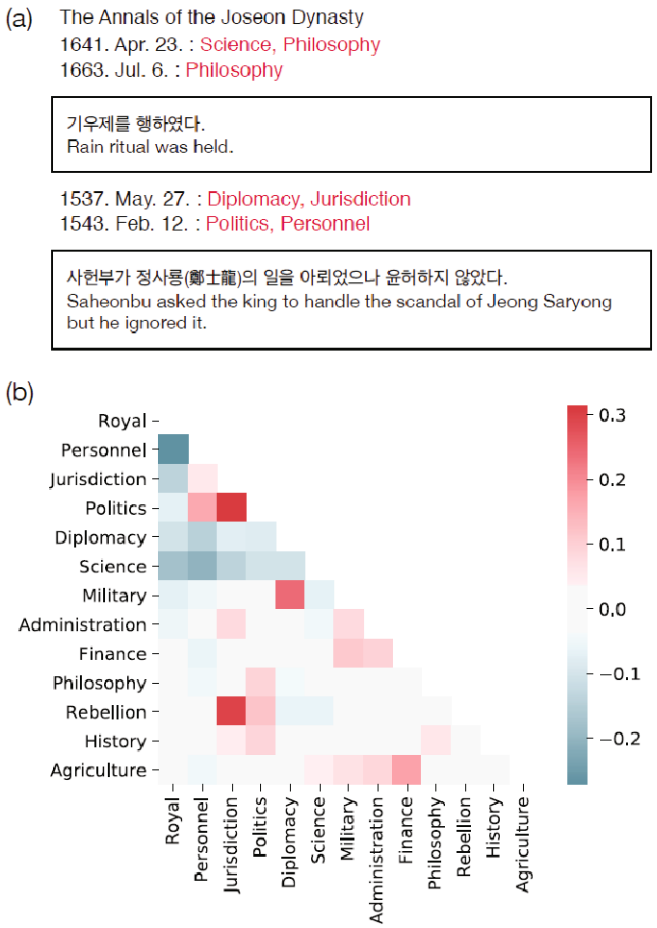

Fig. 1. (Color online) (a) Examples of articles with the same content but different categorization by year. (b) Correlation matrix between the frequencies of 13 different categories in the AJD. The above areas are listed in descending frequency.

Our method of analysis is as follows: We first construct a HAN to read and classify the articles in the AJD by category. We then compare the performance of HAN with there for other existing conventional text classification models and find that the deep learning model (HAN), which understands the context of the whole document, performs better than the word-frequency-based models that do not consider contextual elements. Next, to confirm the performance of our model in a real-world application, we apply it to an uncategorized historical document, *The Daily Records of Royal Secretariat of Joseon*

Dynasty [15]. Finally, by extracting words as document features that play major roles in the classifying process and using the extent of attention for each word, we study how the representative words for each article category in the AJD vary over time. As a result, we reveal that the categories are divided into two groups: one that hardly changes over time, and one that rapidly changes over time. We expect our model will accelerate the massive classification work of both historians and translators by presenting them with the significant sentences and words of large-scale texts as determined by the model in the classification process.

II. DATA

The *Annals of the Joseon Dynasty* (AJD) is a historical record written in classical Chinese written in chronological order and covers 472 years (1392–1863) including the reigns of 25 kings. Notably, this document deals with a longer period than any other historical document in the world. The National Institute of Korean History runs a web service that can be used to browse the Korean translations and original texts directly.

A total of 380,009 articles, which are labeled by year, month, and day, can be found in the AJD. Historians have classified the articles into 40 categories (Royal, Political, Administration, Personnel, Jurisdiction, Military, Diplomacy, Rebellion, People, Finance, Monetary, Price, Commerce, Trade, Transportation, Measures, Agriculture, Fishing Industry, Mining Industry, Industry, Construction, Family, Census, Social Status, Country, Clothing Habits, Dietary Life, Housing Life, Ethics, Custom, Relief Work, Health, Philosophy, History, Science, Medicine, Language, Arts, Education, Publication). An article is multi-labeled if it contains multiple relevant categories; for example, the second article in Fig. 1(a) has two labels, Diplomacy and Jurisdiction.

III. METHODS

The task of classifying datasets that have more than one label is called multi-label classification. In this study, we build an ML model that classifies articles by using the sentences and words of the articles as features. Particularly, the HAN makes performing multi-label classification possible, as this model is trained based on the hierarchical structure of a given text. One additional advantage of the HAN is that it shows the extent to which it focuses on particular sentences and words during classification. The process of building the model is as follows:

1. Data Preprocessing

The 40 categories of the AJD appear inhomogeneously. To mitigate this imbalanced effect on training, we consider only 13 categories with at least 10,000 appearances: Royal, Politics, Administration, Personnel, Jurisdiction, History, Military, Finance, Diplomacy, Agriculture, Philosophy, Rebellion, and Science. These 13 categories account for 87.23% of the total number of categories appearing in the articles of the AJD. For a basic statistic, we measured the Pearson correlations between categories; Fig. 1(b) illustrates the correlation between each category as expressed by using a heat map. While most are weakly correlated, some categories are relatively strongly correlated. For example, in the case of Royal and Personnel, a negative correlation exists, but in the case of Rebellion and Jurisdiction, a positive correlation exists.

Working with the official Korean translations of the AJD, we use a Korean morphological analyzer tool [2] to separate the text data into sentences and words. A total of 77,266,147 words are extracted from the morpheme analyzer, and only the words appearing at least 5 times are taken to build the vocabulary dictionary to a total size of 69,669 words. The average number of sentences per article is 25.8, and the average number of words per article is 170.

In order to employ the widely used cross-validation method to evaluate the performance of the ML model, we divided the whole dataset into a training set, a validation set, and a test set in respective ratios of 0.7, 0.1, and 0.2. In the validation set, the hyperparameters of the following models are set to achieve the best performance.

2. Hierarchical Attention Network

In this subsection, we introduce the components of the HAN in detail. The overall architecture is illustrated in Fig. 2.

Word Embedding: We use the `gensim` [16] package to train the word-embedding parameter of Word2Vec [17]. In order to convert the entire dictionary into a vector space of 200 dimensions, we need a matrix of $\mathbb{R}^{69,669\times 200}$, which is called the word-embedding parameter $W^{(e)}$. The matrix $W^{(e)}$ converts the words to real-space input vectors of the HAN. When training the model, the word-embedding parameter is updated as well.

Gated Recurrent Unit: The gated recurrent unit (GRU) [18] uses a gate mechanism that preserves the state of a sequence without using memory cells. The GRU consists of an update gate (z_t) and a reset gate (r_t), which are found by using

$$z_t = \sigma\left(W^{(z)}x_t + U^{(z)}h_{t-1} + b^{(z)}\right), \quad (1)$$

$$r_t = \sigma\left(W^{(r)}x_t + U^{(r)}h_{t-1} + b^{(r)}\right), \quad (2)$$

where $\{W^{(z)}, U^{(z)}, b^{(z)}, W^{(r)}, U^{(r)}, b^{(r)}\}$ is a trainable parameter set of the GRU that is updated in the learning process. The input vector x_t at current time t and the hidden state vector h_{t-1} at a previous time $t-1$ are linearly combined and passed into the sigmoid activation function $\sigma(x) = 1/(1+e^{-x})$. The GRU calculates the new hidden state using

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t, \tag{3}$$

where $\odot$ is the Hadamard product, h_{t-1} is past information, and $\tilde{h}_t$ is new information to remember. Equation (3) is the linear interpolation between h_{t-1} and $\tilde{h}_t$. Therefore, the update gate z_t determines the amount of past information and new information to be preserved. The newly added information $\tilde{h}_t$ can be obtained using

$$\tilde{h}_t = \tanh\left(W^{(h)} x_t + r_t \odot \left(U^{(h)} h_{t-1}\right) + b^{(h)}\right). \tag{4}$$

The reset gate r_t determines how old information h_{t-1} is to be used when calculating new information. If r_t is zero, the previous information is completely neglected.

Word Encoder: An i-th sentence, consisting of a total of T words, is represented as w_{it}, $t = 1, \ldots, T$. Then, this is translated into a vector by word embedding through $x_{ij} = W^{(e)} w_{ij}$. Word vectors can be represented as a single sentence vector by using a bidirectional GRU [18]:

$$\begin{aligned} x_{ij} &= W^{(e)} w_{ij}, \\ \overrightarrow{h}_{it} &= \overrightarrow{GRU}(x_{it}), \\ \overleftarrow{h}_{it} &= \overleftarrow{GRU}(x_{i(T-t+1)}),\ t = 1, \ldots T. \end{aligned} \tag{5}$$

Here, $h_{it} = [\overrightarrow{h}_{it}, \overleftarrow{h}_{it}]$ is expressed by the forward GRU and the backward GRU, and h_{it} is the information that summarizes the whole context of the i-th sentence w_{it}.

Attention Mechanism [18,19]: Generally, the words in a given sentence do not evenly imply the meaning of the sentence. The attention mechanism is a practical method to leverage this characteristic. The attention a_{it} can be calculated as follows:

$$\begin{aligned} u_{it} &= \tanh\left(W^{(w)} h_{it} + b^{(w)}\right), \\ a_{it} &= \frac{\exp\left(u_{it}^T u^{(w)}\right)}{\sum_t \exp\left(u_{it}^T u^{(w)}\right)}, \end{aligned} \tag{6}$$

where $\{W^{(w)}, b^{(w)}, u^{(w)}\}$ is a trainable parameter set, and $u^{(w)}$, called the context vector, is the basis of word importance. Note that a_{it} is the weight of the word w_{it} in the sentence, and that the sentence vector s_i can be expressed as

$$s_i = \sum_t a_{it} h_{it}. \tag{7}$$

Sentence Encoder: For the sentence vector sequence $\{s_1, s_2, \ldots, s_L\}$, the document vector d can be defined in a similar way. Likewise, encoding each sentence by using the bidirectional GRU can be done as follows:

$$\begin{aligned} \overrightarrow{h}_i &= \overrightarrow{GRU}(s_i), \\ \overleftarrow{h}_i &= \overleftarrow{GRU}(s_{L-i+1}), \\ h_i &= [\overrightarrow{h}_i, \overleftarrow{h}_i],\ i = 1, \ldots, L. \end{aligned} \tag{8}$$

The vector h_i is the information that sentence i computes considering the entire context of the document. The attention vector a_i and the document vector d are written in the same form as follows, where $\{W^{(s)}, b^{(s)}, u^{(s)}\}$ is a trainable parameter set:

$$\begin{aligned} u_i &= \tanh\left(W^{(s)} h_{it} + b^{(s)}\right), \\ a_i &= \frac{\exp\left(u_i^T u^{(s)}\right)}{\sum_t \exp\left(u_i^T u^{(s)}\right)}, \\ d &= \sum_i a_i h_i. \end{aligned} \tag{9}$$

Multi-Label Classification: The document vector d can be used as an input to the classification model as a feature vector. To combine multiple logistic regression models among the methods typically used in multi-label classification, we adopt

$$y_c = \sigma\left(W_c^{(d)} d + b_c^{(d)}\right),\ c = 1, \ldots, C, \tag{10}$$

where y is the predicted label vector and y_c represents the probability that the document would have category c. The total number of categories is C; in this study, $C = 13$. $\{W_c^{(d)}, b_c^{(d)} | c = 1, \ldots, C\}$ is a trainable parameter set. We use a negative log likelihood as the loss function, which is minimized in training as

$$L = -\sum_{n=1}^{N} \sum_{c=1}^{C} \left[t_c^n \log y_c^n + (1 - t_c^n) \log\left(1 - y_c^n\right)\right], \tag{11}$$

where t^n is the target label vector of the n-th document. Here, t_c^n represents whether document n has catergory c; for instance, if it has category c then $t_c^n = 1$; otherwise, $t_c^n = 0$. The size of the total training dataset is N, and n represents the index of the document.

Hyperparameters: The hyperparameters are set as follows: The dimension of the hidden state vector in the GRU is set to 100, and in this case, it has 100-dimensional vectors for forward and backward directions. The word context vector $u^{(w)}$ and the sentence context vector $u^{(s)}$ have 200 dimensions and are randomly initialized before the start of training. We select Adam [20] as our optimizer and set the mini-batch size to 64 and the learning rate to 0.001. Under the above settings, we use `pytorch` [21] for model training and evaluation.

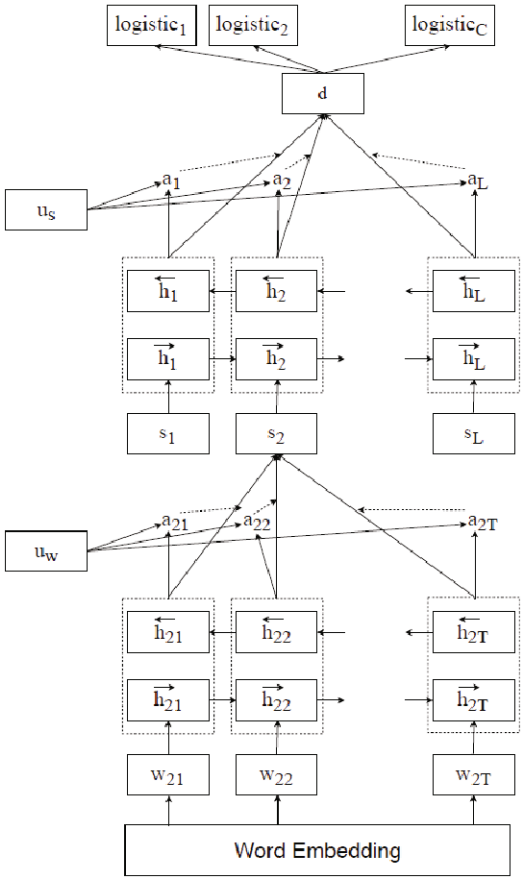

Fig. 2. Architecture of our Hierarchical Attention Network with multiple logistic classifiers for multi-label classification.

3. Baselines

Bag-of-means: Typically, a technique is chosen to express a document as a vector to be used as the feature in a text classification model. One such technique is called the bag-of-means [17], which is the average of all word vectors appearing in a document and can be found by using

$$\text{Bag-of-means} = \frac{1}{\sum_i T_i} \sum_{i=1}^{L} \sum_{j=1}^{T_i} W^{(e)} w_{ij}, \qquad (12)$$

where L is the number of sentences and T_i is the number of words in the i-th sentence. In this study, a total of 13 binary classifiers are trained for each category by using bag-of-means as a feature.

***N*-gram + TF-IDF**: An N-gram [3] is a continuous sequence of N words. It is a typical method that converts a document to a feature vector whose elements represent the frequency of the N-gram. Another useful feature is called TF-IDF [22], short for term frequency-inverse document frequency, which indicates how important a word is in a document of a corpus. In this study, we define the feature of a document by using the TF-IDFs of N-grams composed of $N < 4$. The classifier uses a logistic regression as above.

Naïve Bayes: The Naïve Bayesian classifier [23] is a probabilistic classifier using the Bayesian theorem with independence between features. As the most popular text classification model, it is widely employed as a binary classifier that uses word frequency as a feature vector.

The above three baselines are trained with `scikit-learn` [24]. We use a grid search method to find the best fit hyperparameters.

IV. RESULTS

1. Performance

All models classify the document n to be labeled as c when the output of the model y_c^n is greater than 0.5. Accuracy in multi-label classification is defined as $|\mathbf{T} \cap \mathbf{P}|/|\mathbf{T} \cup \mathbf{P}|$, where $\mathbf{T}$ is the set of true labels and $\mathbf{P}$ is the set of predicted labels. The Hamming loss is defined as the number of wrong labels $|\mathbf{T} \oplus \mathbf{P}|$, where $\oplus$ is the bitwise XOR operation divided by the total number of labels NC, where N is the number of articles and C is the number of categories. Macro F1 is the average of the F1 scores of each class, whereas Micro F1 computes the global F1 score by counting the total true positives, false negatives, and false positives. Macro F1 does not consider label imbalance, but Micro F1 does [25]. Table 1 compares the multi-label classification performance of the models. As can be seen, HAN shows the best performance throughout all performance metrics. In Table 1, 'HAN without names' means that all the people's names in the training set are replaced with 'someone', as discussed later. The performance of this model is slightly lower than that of the HAN.

2. Prediction and Attention

Figure 3 shows the HAN prediction and true labels of an article from the AJD. As previously discussed, the HAN has an advantage in that it shows which sentences and words are important during classification tasks. The attention values of the words and sentences are expressed in red and blue, respectively; the higher the opacity, the greater the attention of the word (sentence) is in the sentence (article). In the case of Fig. 3, the first, second, and fifth sentences are the most noticeable, and the words '산송 (Mountain litigation)', '판결 (Verdict)', '고신 (Certification)' and '빼앗겼다 (Taken away)' in the first sentence are judged to be the most important keywords. Considering the translation of the first sentence, "The executive certification of Lee Minjeok ... was taken away because he made the wrong verdict in the mountain litigation," This article can be, indeed, confirmed to be related to Personnel and Jurisdiction.

Table 1. Multi-label classification results. Accuracy, Hamming loss, micro F1, and macro F1 are the evaluation metrics. The best performances are in bold.

Metrics	HAN	HAN without names	Bag-of-means	N-gram + TF-IDF	Naïve Bayes
Accuracy	**71.0%**	70.5%	66.9%	64.8%	63.3%
Hamming loss	**0.044**	0.046	0.0522	0.0519	0.0590
Micro F1	**0.83**	0.82	0.80	0.79	0.78
Macro F1	**0.75**	0.74	0.71	0.65	0.64

Table 2. Representative words in the Science and Rebellion categories of the AJD from 1500 to 1550 and from 1650 to 1700.

	Science	Rebellion
1500 ∼ 1550	가뭄 drought, 간방 northeast, 건방 northwest, 경천 worship of heaven, 곤방 southwest, 관상감 meteorological administration, 구름 cloud, 기우제 rain ritual, 꼬리 tail, 나타났다 appear, 날씨 weather, 낮 noon, 내렸다 fall, 대낮 high noon, 묘방 east, 무지개 rainbow, 미시 1–3 PM, 밤 night, 백기 white flag, 번개 lightning, 보였다 observed, 불 fire, 비 rain, 빛깔 color, 서리 frost, 손방 southeast, 양 sheep, 오방 south, 오시 11 AM to 1 PM, 우박 hail, 운기 cloud shape, 유성 meteor, 일식 solar eclipse, 재변 disaster, 지진 earthquake, 진동 oscillation, 천둥 thunder, 충해 crop damage due to vermin, 태백 Venus, 한재 drought disaster, 햇무리 parhelion, 현 chord, 혜성 comet, 흰 white	경빈박씨(敬嬪朴氏) Gyeong Binbakssi, 국모 queen, 김극핍(金克) Kim Geukpib, 김억제(金億濟) Kim Eokje, 김의(金毅) Kim Eui, 김일손(金馹孫) Kim Ilson, 김형(金泂) Kim Hyung, 나세찬(羅世纘) Na Sechan, 나옥수(羅玉守) Na Oksoo, 노영정(盧永貞) No Youngjung, 서증(徐) Seo Jeung, 선한(宣漢) Sun Han, 신문 interrogate, 심사순(沈思順) Sim Sasoon, 오국동(吳國同) Oh Gukdong, 완천정(完川正) Wan Cheonjeong, 우안민(禹安民) Woo Anmin, 유경인(柳敬仁) Yoo Kyungin, 유세창(柳世昌) Yoo Sechang, 윤탕빙(尹湯聘) Yoon Tangbing, 이과(李顆) Lee Kwa, 이연(李淵) Lee Yeon, 이장종(李長宗) Lee Jangjong, 이줄(李茁) Lee Jool, 정막개(鄭莫介) Lee Makgae, 지윤해(池允海) Ji Yunhae, 추대 have a person as head, 형장 the place of execution, 홍여(洪礪) Hong Yeo, 홍우(洪祐) Hong Woo
1650 ∼ 1700	각성 α Virginis, 간방 northeast, 곤방 southwest, 금성 Venus, 기우제 rain ritual, 꼬리 tail, 나타났다 appear, 남두 Sagittarius, 내렸다 fall, 눈 snow, 달무리 moon halo, 달 moon, 두성 ϕ Sagittarii, 목성 Jupiter, 묘성 Pleiades, 무지개 rainbow, 미시 1–3 PM, 밤 night, 번개 lightning, 범 tiger, 북두성 Big Dipper, 불빛 light, 비 rain, 빛깔 color, 사지 paper-making office, 서원 auditorium, 손방 southeast, 안개 fog, 우박 hail, 유성 meteor, 일식 solar eclipse, 재이 natural calamity, 정성 μ Geminorum, 주먹 fist, 지진 earthquake, 진시 7:30 to 8:30 AM, 천둥 thunder, 태백성 Venus, 토성 Saturn, 해일 tsunami, 햇무리 parhelion, 화성 Mars, 흘러갔다 flow	강오장(姜五章) Kang Ohjang, 국청 권부(權扶) Kwon Bu, 김성(金成) Kim Sung, 김시정(金時鼎) Kim Sijeong, 김일경(金一鏡) Kim Ilgyeong, 김천추(金天樞) Kim Cheonchu, 김춘택(金春澤) Kim Choontaek, 민언량(閔彦良) Min Eonryang, 민진원(閔鎭遠) Min Jinwon, 박도창(朴道昌) Park Dochang, 심유현(沈維賢) Sim Yoohyun, 업동(業同) Eob Dong, 역적 traitor, 유봉휘(柳鳳輝) Yoo Bonghwi, 유성추(柳星樞) Yoo Sungchu, 이감(李) Lee Kam, 이선행(李善行) Lee Seonhaeng, 이의연(李義淵) Lee Uiyeon, 이인관(李仁寬) Lee Ingwan, 임서호(任瑞虎) Im Seoho, 장희재(張希載) Jang Heejae, 정국 interrogate, 조송(趙松) Jo Song, 죄인 prisoner, 주노미(周老味) Joo Nomi, 친국 interrogate by the king, 토죄 scold severely for committing a crime

3. Application

As an application, in this section, we test whether our HAN, trained with the AJD, also works well for classifying *The Daily Records of Royal Secretariat of Joseon Dynasty* (SJW), which is a historical text comprising daily records written during the Joseon Dynasty by an organization called Seungjongwon [15]. Registered as a Korean National Treasure and also as a UNESCO World Record Heritage, this text has yet to be officially classified. Also written in classical Chinese, the daily records of the SJW include administrative tasks and orders of the Joseon kings, and even weather and astronomical phenomena. The editors of the AJD used the SJW as a reference material when compiling the AJD. Comparing the two, while the SJW covers a shorter period (1623–1910), its total size is about 4.5 times that of the AJD (242 million compared to 54 million characters). Due to its vast size, only 22% of the SJW has been translated to Korean as of December 2018.

We present the results of the classification of two SJW articles by using the HAN. We randomly selected a number of articles of various topics in SJW and tested our model to classify them. Figure 4 shows the two example articles and the classification results with the attention values of the sentences and words. One of the main advantages of the model is that the obtained confidence of each classification lies between 0 and 1 when using the model's final logistic regression, which is informative for human classifiers in real world situations. The article in Fig. 4(a) is categorized as Science (confidence = 0.9998), and a high level of attention is given to the words 'Red' (relating to the sun), 'Parhelion', and 'Meteor'. The ar-

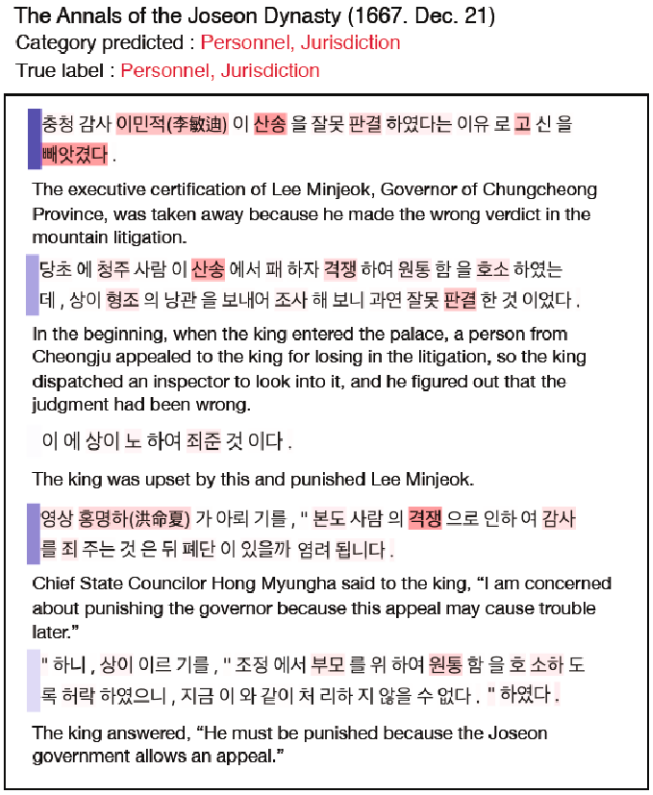

Fig. 3. (Color online) Prediction obtained by HAN and true labels. English translations are presented Attention values of sentences (blue) and words (red) are represented by opacity.

ticle in Fig. 4(b) contains a call for the dismissal of a person by Saganwon, a government institution that gave advice to the king, and reports on issues in the barracks. Our trained model classified this article into the following categories: Politics (confidence = 0.9863), Personnel (0.9950), Jurisdiction (0.9848), and Military (0.9612), which are consistent with the actual descriptions of the article.

4. Temporal Change of Representative Words

In this section, we try to understand how the historical document evolves in time for each category, and we find two different groups of categories. To analyze temporal changes in historical categories, we extract the representative words that have the highest total attention for each AJD category. Table 2 lists the representative words from two sample categories, Science and Rebellion, during two 50-year time periods, 1500–1550 and 1650–1700. In the case of Science, the words '기우제 (Rain ritual)', '지진 (Earthquake)', '태백성 (Venus)', and '햇무리 (Parhelion)' had high attention throughout the centuries. On the other hand, in the case of Rebellion, the words in the early and the late periods had little overlap.

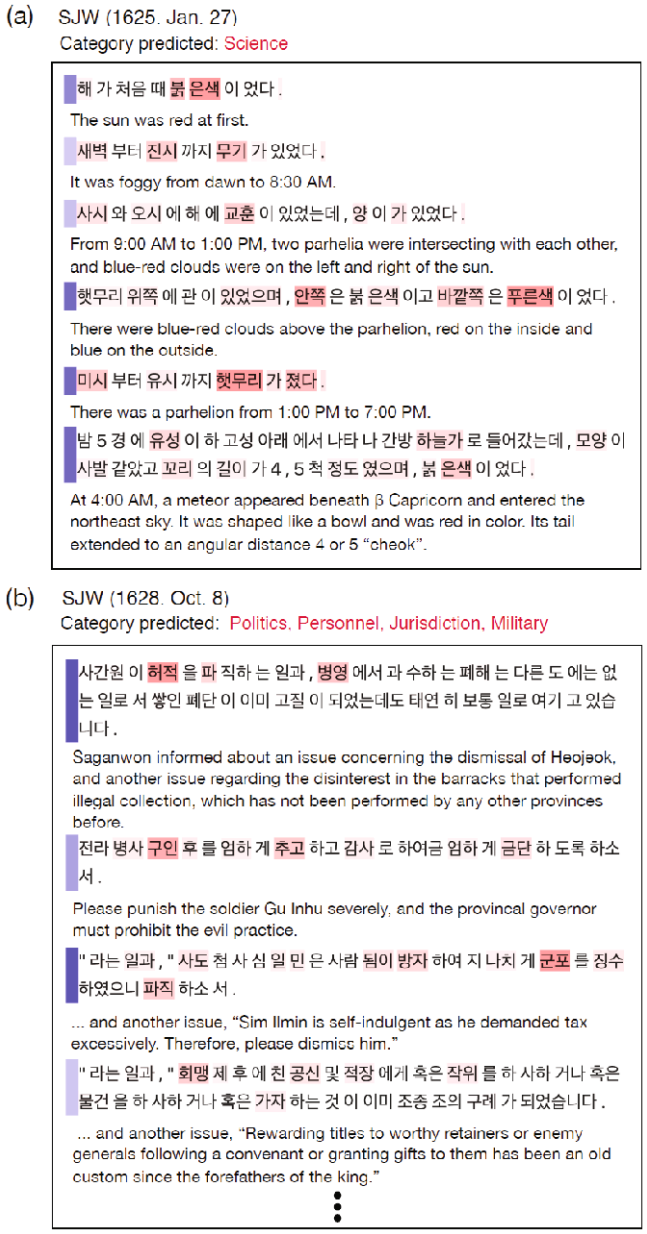

Fig. 4. (Color online) Category prediction using HAN on two sample articles from *The Daily Records of Royal Secretariat of Joseon Dynasty* (SJW). (a) Content of an SJW article from January 27, 1625, and predicted category. (b) Content of an SJW article from October 8, 1628, and predicted categories. Attention values of sentences (blue) and words (red) are represented by opacity.

Table 2 shows that the words in the Rebellion category are mostly the names of the people who were actually involved in the rebellions of that time period and, thus, do not often appear again outside of their time period. Along these lines, we further found that the categories of Royal, Military, Diplomacy, Finance, Agriculture, and Science have relatively higher attention on verbs and nouns than on people's names [26]. Although people's

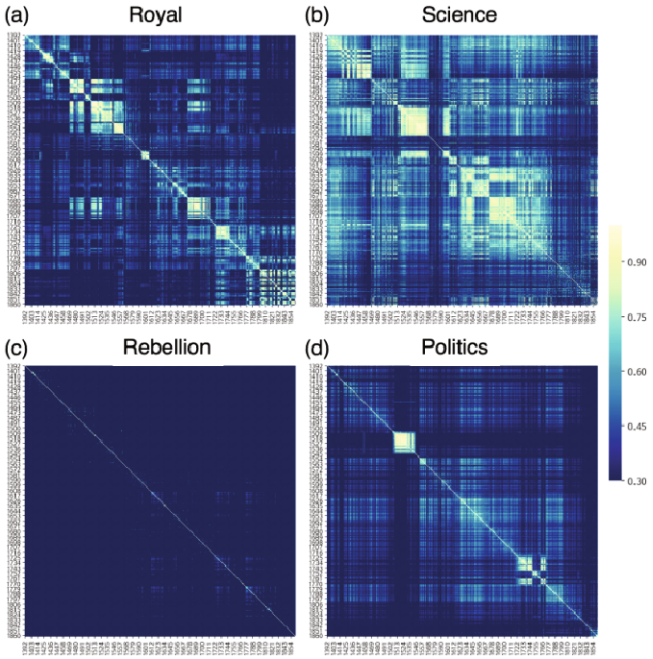

Fig. 5. (Color online) Pearson correlation between different years from 1392 to 1863 for (a) Royal, (b) Science, (c) Rebellion, and (d) Politics categories of the AJD. Correlation maps of all categories can be found in Fig. S1 in the Appendix.

names often appeared as representative words, as seen in Table 1, the 'HAN without names' model shows that this does not have a big impact on performance.

The following methodology was used to examine the changes in the attention of words over time: First, we select the top 5,000 words with the highest attention values from the vocabulary dictionary. Next, we define an attention vector, in which each i-th element represents the total attention summation of the i-th word, for each year of the AJD. In other words, attention vectors were obtained for each year from 1392 to 1863. Figure 5 shows the Pearson correlation between different years in four categories: (a) Royal, (b) Science, (c) Rebellion, and (d) Politics. In the case of Royal, a strong correlation is found for specific intervals, which is consistent with the reigns of each king (Fig. 5(a)). In the Science category, a strong correlation is found across all time periods, indicating that scientific terms were used with no significant change over time (Fig. 5(b)). In the Rebellion category, no temporal correlation is found (Fig. 5(c)), and in the case of Politics, a strong correlation is found in particular time periods (Fig. 5(d)). For example, the words of the sentences have a very similar attention distribution for about 50 years from about 1500 to 1550.

V. CONCLUSIONS

Quantitative analyses of digitized historical documents has been proliferating in recent years. Text classification is of particular importance to quantitative historical analyses because it helps researchers search literature efficiently and determine the important subjects of a particular age. However, because extensive historical records are analyzed jointly by various historians, consistent judgments are not always applied to similar documents. This problem of inconsistency also exists in a large-scale Korean historical document, the *Annals of the Joseon Dynasty* (AJD), as well as in other historical records, and is considered a critical issue in the quantitative analysis of history. In this study, we trained a deep-learning model to classify the article categories of the AJD. The Hierarchical Attention Network (HAN) showed better performance than previous models by taking into account the entire context of the work rather than simply classifying the categories by word frequency. Analyzing sets of words with high attention that describe each category over time, we found two groups of categories. In some categories, terms directly related to the category showed the highest attention values (such as 'Earthquake' or 'Rain' in the Science category) whereas in other categories such as Rebellion, people's names showed the highest attention values. This is a characteristic of machine-learning algorithms compared to human experts because humans usually decide the category of an article based on the contents rather than individual names.

We further investigated the changes in the attention of words over time. In some categories, strong correlations were found for specific intervals, which is consistent with the reigns of kings or the consistent use of terms over time (such as Royal and Science) whereas in other categories, little temporal correlation was found (such as Politics and Rebellion). Finally, we checked whether the model trained in this study could contribute to the consistent classification of other uncategorized historical documents, such as the *The Daily Records of Royal Secretariat of Joseon Dynasty* (SJW). Via application of our HAN model to the SJW, we determined that plausible classification can be achieved. We expect that such textual analysis using deep learning will contribute to the consistent classification process and further afford a quantitative understanding of sociocultural phenomena.

ACKNOWLEDGMENTS

This work was supported by the National Research Foundation of Korea (Grant No. 2017R1A2B3006930).

APPENDIX

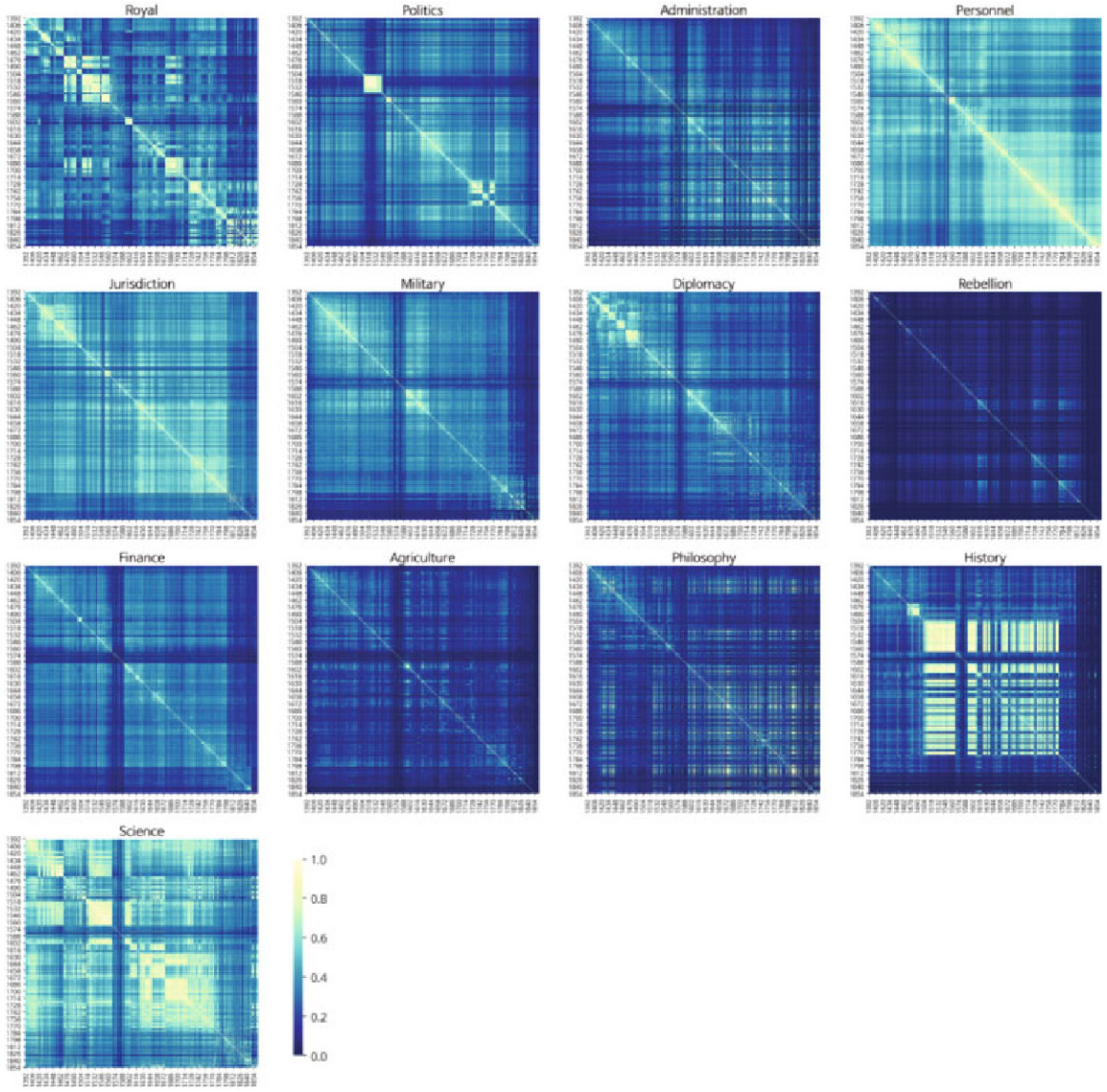

Fig. S1. (Color online) Pearson correlation maps between different years from 1392 to 1863 for the 13 chosen categories of the AJD.

REFERENCES

[1] D. J. Hopkins and G. King, Am. J. Political Sci. **54**, 229 (2010).
[2] J. Grimmer and B. M. Stewart, Polit. Anal. **21**, 267 (2013).
[3] J. B. Michel *et al.*, Science **331**, 176 (2011).
[4] S. Klingenstein, T. Hitchcock and S. DeDeo, Proc. Natl. Acad. Sci. U.S.A. **111**, 9419 (2014).
[5] S. Hochreiter and J. Schmidhuber, Neural Comput. **9**, 1735 (1997).
[6] Y. Wu *et al.*, arXiv:1609.08144.
[7] D. Tang *et al.*, in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (Baltimore, Maryland, USA, June 23–25, 2014), Vol. 1, pp. 1555–1565.
[8] Y. Kim, arXiv:1408.5882.
[9] X. Zhang, J. Zhao and Y. LeCun, in *Advances in Neural Information Processing Systems* (Montreal, Canada, December 7–12, 2015), pp. 649–657.
[10] Z. Yang *et al.*, in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (San Diego, CA, USA, June 12–17, 2016), pp. 1480–1489.

[11] B. Lee, D. Kim, D. Kim and H. Jeong, New Phys.: Sae Mulli **66**, 502 (2016).
[12] J. Bak and A. Oh, in *Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences and Humanities (LaTeCH)* (Beijing, China, July 30, 2015), pp. 10–14.
[13] J. Bak and A. Oh, in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (Brussels, Belgium, October 31–November 4, 2018), pp. 956–961.
[14] The Annals of the Joseon Dynasty, http://sillok.history.go.kr.
[15] The Daily Records of Royal Secretariat of Joseon Dynasty, http://sjw.history.go.kr.
[16] R. Rehurek and P. Sojka, in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, (Valletta, Malta, May 22, 2010), pp. 45–50.
[17] T. Mikolov, K. Chen, G. Corrado and J. Dean, arXiv:1301.3781.
[18] D. Bahdanau, K. Cho and Y. Bengio, arXiv:1409.0473.
[19] K. Xu *et al.*, in *International Conference on Machine Learning* (Lille, France, July 6–11, 2015), pp. 2048–2057.
[20] D. P. Kingma and J. Ba, arXiv:1412.6980.
[21] A. Paszke *et al.*, in *31st Conference on Neural Information Processing Systems* (Long Beach, CA, USA, December 4–9, 2017).
[22] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval* (McGraw-Hill, New York, NY, USA, 1983).
[23] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach* (Pearson Education Limited, Malaysia, 2016).
[24] F. Pedregosa *et al.*, J. Mach. Learn. Res. **12**, 2825 (2011).
[25] G. Tsoumakas and I. Katakis, Int. J. Data Warehous. Min. **3**, 1 (2007).
[26] The ratio of people's names to verbs and nouns in each category is as follows; Royal 0.11, Military 0.13, Diplomacy 0.18, Finance 0.10, Agriculture 0.10, Science 0.01, Politics 0.58, Administration 0.30, Personnel 0.60, Jurisdiction 0.51, Rebellion 0.60, Philosophy 0.25 and History 0.42.