Journal of Library and Information Studies 20:1 (June 2022) pp.1-30
https://doi.org/10.6182/jlis.202206_20(1).001

Practicability of Ensemble Artificial Neural Network Models for a Classification Task: An Optimal Approach for Reproducing Classification Practices of Health Consumers Generated Text on Social Media

Sukjin You[1], Min Sook Park[2], Soohyung Joo[3]

Abstract

This paper reports the classification accuracy of artificial neural network (ANN) models in reproducing health consumers' classification practices in social media. Social media have driven the growth of unstructured text data across domains including health, which motivates researchers to reconsider the epistemological approach to automated classification. This study compared the performance of several types of ANN models and ensemble models based on classification results and the integration of multiple ANN structures. To train these models, two dictionaries were employed: health consumers' terms extracted from questions and answers in the health categories of Yahoo!Answers and MeSH terms. All three types of individual classifiers demonstrated accuracies of around 90%. In particular, the fully connected ANN with two layers produced relatively higher classification performances than a convolutional neural network and long short-term memory. Ensemble models based on classification results outperformed not only the ensemble models based on the integration of heterogeneous ANN structures but also individual deep-learning models. The combination of questions and best answers were found to be most effective as a training dataset to build an accurate prediction model. The findings suggest that ANN models can be an effective assistive tool in classifying online health resources generated by health consumers in natural language.

Keywords: Automated Classification; Deep Learning; Artificial Neural Network; Ensemble Classification Model; Knowledge Organization

1. Introduction

The popularity of social web technologies has paved the way for information users to participate in information creation and consumption using social platforms (e.g., social Q&A sites, blogs, and social network sites) across different domains including health. These activities have accumulated as online resources, reflecting diverse experiences and best practices from health consumers' perspectives (Andersen & Söderqvist, 2012; Kamel Boulos & Wheeler, 2007; O'Reilly, 2007). However, the sheer amount and unstructured nature of the resources place limitations on the ability to organize information and have resulted in difficulty locating relevant information (Cline & Haynes, 2001).

[1,2] School of Information Studies, University of Wisconsin at Milwaukee, Wisconsin, USA
[3] School of Information Science, University of Kentucky, Lexington, Kentucky, USA
* Corresponding Author: Min Sook Park, E-mail: minsook@uwm.edu

Classification has been discussed as necessary to achieve the web's full potential as an information sphere, increasing access for web users to relevant information (Pierre, 2001). The initial applications of conventional knowledge organization systems (KOSs; e.g., directory, metadata, and indexing) to web resource classification were useful, but showed limited abilities in coping with large-scale resources. Leveraging data analytic techniques for classification has been discussed as a promising alternative to build and maintain KOSs that are useful, flexible, and practical (Ibekwe-Sanjuan & Bowker, 2017; Shiri, 2013), and more applicable across domains (Pierre, 2001; Weller, 2010).

Besides, organization of user-generated resources online requires a bottom-up approach, whereas conventional KOSs largely remain in the purview of experts rather than user perspectives (Greenberg, 2003; Ibekwe-Sanjuan & Bowker, 2017; Pierre, 2001; Weller, 2010). Health consumers in social media are known to use their own vocabulary to describe their health issues, which is different from established controlled vocabularies (Kim, 2013; Messai et al., 2010; Poikonen & Vakkari, 2009). This vocabulary gap is also known to hinder health consumers' access to relevant resources on the web (Gross & Taylor, 2005; Seedorff et al., 2013; Smith & Wicks, 2008). Because the primary common goal of classification is linking users to knowledge resources to satisfy the users' needs (Svenonius, 2000), sufficient commonality is required between the concepts expressed in a KOS and objects in the real world to which that concept refers (Abbas, 2010; Weller, 2010). In this regard, organization methods for health consumer-generated resources need to not only handle large-scale data but also reflect users' own concepts and practices in a domain.

Machine learning has been a dominant approach to automated classification since the early 1990s (Sebastiani, 2002). Of its various manifestations, artificial neural networks (ANNs) present strong potential to classify socially generated health resources, effectively reproducing human classification practices (Khan et al., 2001; Kim, 2014). The selection of an adequate machine learning model for specific application contexts and having the right kind and amount of data to train the algorithms is a premise to consider in reproducing human classification practices (Hartmann et al., 2019). However, previous major endeavors have used good-quality texts with high homogeneity. As an effort to identify effective classification systems in the context of consumer-generated health information, this study evaluates ANN models, based classifiers with different training data and configurations. The detailed research objectives are as follows:

The first objective was to evaluate the ability of ANN models, and explore the applicability of ANN models and ensemble models based on multiple classifiers by comparing the classification accuracy of those models. The designated ANN models were a deep neural network (DNN) based on two fully connected layers, a convolutional neural network (CNN), and long short-term memory (LSTM), along with ensemble deep-learning models based on the multiple ANN models. The second objective was to identify optimized training datasets and features that can maximize the performance of ANNs in classifying resources in social media. The performance of

machine learning models often relies on the size and quality of the training dataset and features. Accordingly, the noisy nature of text in social media may affect the performance of ANN classifiers. Thus, there was a need to find a proper dataset that could maximize the ability of classifiers.

2. Background and Related Work

2.1 Background of web resource classification

The advent of social media has led to a need to find optimal methods to organize the large-scale, fast evolving, and unstructured information that social web users generate every second. User-generated information becomes a rich and useful resource across domains, and it reflects users' needs, perspectives, and vocabularies (Greenberg, 2003; Pierre, 2001; Sarasohn-Kahn, 2008). But its size and unstructured nature paradoxically resulted in difficulty with organizing and locating relevant information. Classification has long been discussed as a primary way to increase accessibility to information (Norton, 2010). Efforts to classify resources on the web have featured two distinct approaches: manual classification by human editors and automated classification.

Early efforts to classify web resources mainly involved the application of traditional KOSs developed in library settings. Examples are classification schemes (e.g., Dublin Core) and directories (e.g., Yahoo Directory, Looksmart, Open Directory Project). These early approaches, which heavily relied on human intelligence, have been criticized as unable to keep up with the huge amount of and rapid changes in web content and unstructured nature of user-generated content, thereby making web catalogs quickly obsolete and imprecise (Kamel Boulos & Wheeler, 2007; Khan et al., 2001; Li & Lu, 2008; Peters, 2009; Weller, 2010). Whereas social media reflects health consumers' terminology and needs, developed KOSs use strict rules and reflect the views of experts whose knowledge comes mainly from scientific literature (Greenberg, 2003; Hjørland, 2007; Pierre, 2001). This top-down, prescriptive approach is usually discordant with the dynamic and flexible nature of human conceptual systems (Assefa, 2007).

Automated classification is an alternative and recent approach to organizing web resources. Two major strengths of automated classification are the feasibility of coping with the ever-changing nature of web resources and the ability to meet the needs of different domains with higher precision (Pierre, 2001). Scholars in information studies (IS) have discussed effective alternatives that can yield more descriptive domain KOSs reflecting the current state of knowledge. Among automatic KOS methodologies, text categorization is considered a critical component (Dumais et al., 1998). The knowlege organization (KO) domain has extended to incorporate computational methods for concept extraction for clustering, algorithmic indexing, and automatic classification (Golub, 2019; Golub et al., 2016; Hjørland, 2018; Smiraglia & Cai, 2017). In this sense, automated classification of textual resources has been extensively studied and is now being applied to diverse contexts requiring organization or adoptive document dispatching (Sebastiani, 2002).

Of the available automatic classification methodologies, interest in machine learning approaches has surged since the 1990s. Until the 1980s, knowledge engineering was the most popular approach (Sebastiani, 2002). Knowledge

engineering techniques require domain experts to curate a dictionary, consisting of elaborate word lists and associated labels, a priori, and thereby they are often generic across domains (Hartmann et al., 2019; Sebastiani, 2002). Machine learning bears many advantages over knowledge engineering. The accuracy of text classification using machine learning is comparable to that of human experts while saving considerable manpower (Sebastiani, 2002). Also, machine learning methods use inductive learning in that classifiers are automatically constructed based on observed patterns and assign categories to future documents based on content, thereby requires less expert manpower (Dumais et al., 1998; Lewis & Ringuette, 1994; Sebastiani, 2002). Last, this inductive approach makes machine learning methods flexible in understanding content categories specific to a certain domain.

Although automated classification has advanced mainly in computer science, classification is fundamentally a matter of constructing and maintaining schemes to organize knowledge artifacts, which has been one of the main foci of the information science (IS) discipline. Many classification applications are closely related to or overlap with the major research areas in IS, which is built on philosophical discussions by IS scholars regarding how to rotationally and structurally present humankind's intellectual achievements (Hjørland, 2014; Smiraglia & Lee, 2012; Svenonius, 2000; Tennis, 2008). The science of KO delves into the inquiry of conceptual organization about what is perceived (Hjørland, 2008; Smiraglia, 2015). In that sense, KOS research provides fundamental inputs for areas to manage scaled resources such as text mining, machine learning, and semantic searching (Ibekwe-Sanjuan & Bowker, 2017; Shiri, 2013; Smiraglia & Cai, 2017). A primary philosophical approach underlying KOSs concern not only how known things are represented (i.e., ontology), but also how humans process knowledge (i.e., epistemology). Thus, classification systems reflect a human's self-conscious creation and bear the imprints of their progenitors in the form they take (Dahlberg, 2006; Smiraglia & Lee, 2012), linking humankind's intellectual achievements in a way that satisfies users' needs (Smiraglia & Lee, 2012; Svenonius, 2000). This philosophical approach motivated the current study. The adoption of machine learning, particularly deep learning, outside of the tech sector is at an early stage (McKinsey Global Institute, 2017). Besides, one primary purpose of the automated classification approach is to reproduce human users' classification judgments. Thus, this study explored the applicability of different machine learning models, ANNs and ensemble models, in reproducing health consumers' categorization practices in social media, which reflect their needs, vocabularies, and classification behavior.

2.2 Classification using ANN models and ensembles of ANN models

The selection of adequate methods for specific application contexts is regarded as one of the main challenges to the advancement of machine learning (McKinsey Global Institute, 2017). Previous studies demonstrated that data analytic techniques have also been widely and successfully applied to classification tasks in the health domain (e.g., Agatonovic-Kustrin & Beresford, 2000; Er et al., 2016; Kalantari et al., 2018; Khan et al.,

2001). However, few research studies utilized health consumer-generated content. Besides, they mainly focused on extracting entities from social media for clinicians' use rather than reflecting health consumers' perspectives. To illustrate, Sarker and Gonzalez (2015) devised classifiers based on traditional machine learning techniques to automatically detect adverse drug reactions and medicine use mentioned in social media. Liu et al. (2011) presented automatic question classification methods that distinguished medical questions between health consumers and health care professionals. Zhang et al. (2018) adopted an ANN model to a medical question-answering system in an effort to support finding answers to clinical questions. These studies shed light on adopting a machine learning approach to medical information in social media, but they primarily concentrated on clinicians' use rather than reflecting health consumers' practices and vocabularies. Unlike question classification using synthetic questions or questions asked by health professionals, which are limited to simple entities such as people, places, organizations, and drug names (McRoy et al., 2016), questions from the general public often include complex descriptions, involving diverse medical, daily, and even nonmedical factors to get more tailored answers (Harper et al., 2009; Oh et al., 2016; Zhao & Zhang, 2017).

Machine learning techniques based on ANNs have been popularly employed to solve various classification problems. ANN models are a promising alternative to various conventional classification methods (Zhang, 2000), demonstrating the potential for successful performance in the automated classification of user-generated textual resources. ANNs are computer learning algorithms that are modeled after the way brains process information, allowing a machine to inductively detect patterns and relationships in data it previously processed (Agatonovic-Kustrin & Beresford, 2000; Efron & Hastie, 2016). Several types of neural networks have been designed, such as fully connected ANNs, CNNs, and recurrent neural networks (RNNs), each with its unique strengths. Their flexible structure enables ANNs to perform across different classification tasks (Hartmann et al., 2019). Data-driven self-adoptive abilities that adjust themselves to the data enable ANNs to learn subtle text patterns (Hartmann et al., 2019) and be flexible in modeling real-world complex relationships (Zhang, 2000). In particular, ANNs show superior ability in learning features from heterogeneous unlabeled data (Lin et al., 2014). Due to these advantages, ANN models have been applied to classification tasks, demonstrating high potential to process natural languages (Xu & Rudnicky, 2000). In particular, deep-learning models based on DNNs have achieved remarkable advancements in natural language processing and the categorization of complex patterns (Xu & Rudnicky, 2000).

Notwithstanding their benefits, there are downsides to deep-learning neural networks like ANNs, including high variance in classification models that are trained, even on the same training dataset (Brownlee, 2019). Ensemble models reduce the variance in the final model by employing multiple types of algorithms, data, and intermediate results to predict an outcome, addressing the limitations of individual models. An ensemble approach was applied for web data classification and outperformed individual

machine learning models. For instance, Kim and Cho (2018) proposed an ensemble model that integrated CNN and LSTM to detect anomalous web traffic, and it outperformed other individual machine learning techniques, such as LSTM, CNN, and gated recurrent units. In general, ensemble learning models combine the outputs from multiple models to improve performance by reducing errors from noise, bias, and variance. But methods to build the models vary from averaging results from a group of models to adopting advanced techniques such as bagging (Breiman, 1996), boosting (Freund & Schapire, 1996), and stacking (Wolpert, 1992). These techniques commonly aim to produce stable and robust models, improving prediction results.

Along with choosing the right learning model, selecting the right kind and amount of training data is another important prerequisite for building a good machine learning model (Apté et al., 1994; McCallum & Nigam, 1998; Pierre, 2001). Feature selection for training an ANN model is also considered a prominent part of building an automated classification system. Particularly, in supervised machine learning, pre-classified training data are a key resource. When an inductive process (also known as the learner) automatically builds a classifier based on categories, the learner gleans the characteristics of a set of documents that should have to be classified under those categories (Sebastiani, 2002). Most text categorization methods rely on good-quality documents with high homogeneity (e.g., TREC, Reuters-22578, OSHUMED), especially for training (Jacob, 2014; Pierre, 2001). However, unlike quality texts, web content in social media is heterogeneous and irregular, reflecting the writing of the general public in natural language. The nature of these resources is the main challenge to classifying web resources (Pierre, 2001). Vocabulary gaps between health professionals and the general public is another well-known issue (Zeng & Tse, 2006).

To fill this gap, this study assessed practicability of different types of neural network models—DNN, CNN, LSTM—and ensemble models. In addition, this study also employed different amounts and quality of text-based features. These features were compared with a controlled vocabulary in the medical domain, MeSH, because automatic text classification should be able to support category structures that are general, consistent across individuals, and static (e.g., Dewey Decimal System, MeSH) and those that are more dynamic and customized to individual interests or tasks (Dumais et al., 1998).

3. Method

3.1 Dataset

We collected 1,944,881 posts from Yahoo!Answers (answers.yahoo.com), a social question-and-answer (social Q&A) site, consisting of 289,598 questions and 1,654,283 associated answers. Yahoo!Answers is one of the most popular community-driven Q&A websites, with 114 million visitors as of November 2019 (SimilarWeb, 2019). This social Q&A site enables its users to submit questions and answer questions asked by other users. We selected Yahoo!Answers for this study because: a) the website does not limit the number of words in postings. Its users can elaborate on their health concerns, experience, opinions, and information in postings with sufficient details in questions and answers, thereby

providing opportunities to construct sufficient features and corpora from the postings; b) It also allows its users to generate different textual corpora, which are questions and answers. Of the answers provided by other users, a questioner may select one answer as the "best answer." We viewed these selected answers as user-perceived quality texts that reflect more relevance to the users' needs or the topic of their questions (Kim et al., 2008; Oh et al., 2011); and c) we assumed that the questions assigned to a specific category reflected the users' classification practice because questioners need to select a given category according to their topic.

Python web scraping modules (Calefato et al., 2016) were referred to and revised to collect questions and associated answers posted before June 2018. Questions and all associated answers, including best answerers, were collected from six health categories shown in Table 1 from January to June 6, 2018. These subcategories were chosen to ensure a sufficient dataset size for comparison. To illustrate, the selected categories have 30,000 questions and associated answers of 100,000 or more.

Data were randomly shuffled to control for the interdependence among answers nested in the same category. A testing corpus, reflecting 20% of all collected questions (n = 57,920), was randomly selected and used for all three training document settings. These three experimental corpora were: (a) Corpus Q: a set of randomly selected questions (80%; n = 231,678); (b) Corpus QBA: Corpus Q and associated best answers (n = 457,451); and (c) Corpus QAA: Corpus Q and all associated answers (n = 1,660,188). We also used MeSH terms (U.S. National Library of Medicine, 2018) to compare the performance of neural network models when using controlled vocabulary compared with using natural language from health consumers.

3.2 Text pre-processing and feature selection

To train ANNs, we used the bag-of-words model as the baseline, wherein the occurrence of a single word is used as a feature for training. Then we performed text preprocessing by tokenizing terms and removing stop words (e.g., !, #, and of). Two types of dictionaries were employed in the study as inputs: (a) two sets of user-generated terms extracted from the collected

Table 1. Data Collected from Six Health Categories in Yahoo!Answers

Category	Num. of questions (n)	Num. of answers (Best Answers) (n)	Total (N)
Diet and fitness	34,187	372,833 (31,921)	407,020
Diseases and conditions	38,871	160,803 (28,068)	199,674
General health care	41,104	123,868 (31,370)	164,972
Men's health	39,557	359,749 (34,558)	399,306
Mental health	67,553	393,710 (43,291)	461,263
Optical	68,326	243,320 (56,565)	311,646
Total	289,598	1,654,283 (225,773)	1,943,881

dataset: the 5,000 and 10,000 most frequently observed words; and (b) 21,507 unigram MeSH terms (U.S. National Library of Medicine, 2018). Of the existing 28,472 *n*-gram terms from the "MH" field in the 2018 MeSH descriptor file, 21,507 unique unigram terms were used as the MeSH dictionary. Due to the limited capability of computing, we chose 5,000 terms for the user-generated term dictionary to compare with the MeSH term features. Those 5,000 terms were referenced to generate features for input data, such as word frequency and word embedding vectors. Among those 5,000 user-generated terms, a certain portion of the terms was also found in the MeSH dictionary: 30.8% (n = 1,540) for Corpus Q; 32.32% (n = 1,616) for Corpus QBA; and 33.1% (n = 1,655) for Corpus QAA.

3.3 Classification models

With those features and corpora, we compared different types of classifier models: individual classifiers (i.e., DNN, CNN and LSTM) and two types of ensemble models based on classification results (i.e., health categories of the questions) and heterogeneous structures. Additionally, ensemble models based on classification results were further segmented into two kinds — homogeneous and heterogeneous classifiers. The homogeneous classifiers were constructed using a set of the same type of classifiers whereas the heterogeneous classifiers are a set of classifiers of different type. These two types of ensemble models were compared to explore correlations between diversity and performance of the used classifiers. Lastly, ensemble models that integrated heterogeneous ANN models were compared to the ensemble models based on classification results.

We used a fully connected, supervised network with the backpropagation learning rule, which minimizes the error in prediction by replicating units and weights between neurons through iterative feedforward. To process data and design ANN models, python packages—*genism*, *keras*, and *scikit-learn*—were used. To avoid unnecessary epochs and overfitting, early stopping with a patience number, 5, was applied in the training process. The *softmax* function was used for the output layers, and *rectified linear units* were used as an activation function for hidden or convolution layers. The accuracies of the neural network models using between 5,000 (baseline) and 10,000 unigrams were evaluated.

To evaluate classifiers, accuracy was measured excluding other metrics (e.g., F1 or Cohen's kappa) for early stopping, which was applied to prevent overfitting. Classification results were not notably skewed. In most cases, accuracy was distributed between 80% and 100% for the six health categories. Measured receiver operating characteristic area under the curve (ROAUC) scores were larger (around 99%) than accuracy, which were difficult to compare. Therefore, this measure was not used in the current study.

3.3.1 Individual classifiers

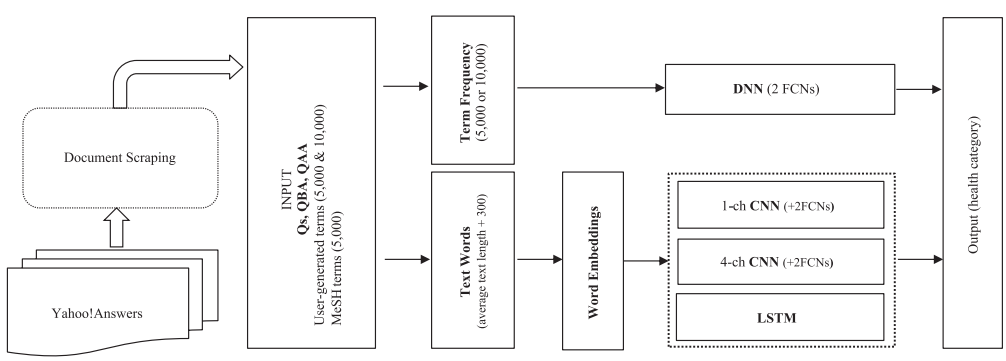
Three types of classifiers based on DNN, CNN, and LSTM were compared. For the DNNs, two hidden layers were set with 128 and 64 nodes. We first used the frequency of individual words as a feature value. Input features for each question and answer were the 5,000 and 10,000 most frequently used words in the dataset. Simply, a sparse matrix for word and frequency was created, despite decreased efficiency in using memory space.

For other classes of deep-learning models, we employed CNN and LSTM. Word embeddings for each word exhibited a high-dimensional structure wherein those vectors can be used as inputs for CNN and LSTM layers. The average text lengths per question and answer in the dataset were less than 100 in three types of datasets, although the text length varied depending on the datasets. For the input for the embedding layer, we used around 400-word sequences for a document (i.e., a question or an answer) (average text length per question or answer in a dataset + 300). Specifically, the number of input word sequences for 5,000 and 10,000 features were 373 and 375 for Corpus Q, 382 and 385 for Corpus QBA, and 355 and 357 for Corpus Q respectively. For short documents, paddings were added after the word indexes for the original text. An embedding layer was added between the input layer and CNN or LSTM layers to represent a word as an embedding vector with real numbers in the 100 dimensions.

CNNs were employed for feature recreation from the original input based on spatial characteristics of the input data. In CNN, input values are convoluted by multiple filters with a specific size of kernels, max-pooled, and then flattened for the input to fully connected networks (FCNs) with two layers. In this study, single-channel and four-channel CNNs were compared for a feature recreation from the original input based on spatial characteristics of the input data. Feature values underwent a transformation process by convolution and max-pooling with kernels and filters. We set up 32 filters with a specific kernel size (e.g., 2, 4, 6, and 8). Recreated feature values were used as inputs for two FCN layers with 128 and 64 nodes.

The LSTM model, which is based on an RNN, was devised to reflect the temporal nature of input data for deep learning. LSTM was adopted in the current study to enhance the influence of previous inputs because LSTM is more effective in maintaining context than RNN by reflecting the outputs from the prior node and other previous nodes. The sequences of input values were considered in updating weights among the nodes, based on an assumption that the output from the prior input node (word) may affect the output of the following input node. Our LSTM model included 128 memory units. Figure 1 illustrates the analytic pipelines of ANN classifiers in the current study.

Figure 1. The Analytic Pipelines of ANN Classifiers



3.3.2 Ensemble classifiers

Two types of models were proposed depending on whether an ensemble model was designed based on (a) classification results or (b) the integrations of different ANN structures. All classifiers for ensemble models were trained on three datasets: Corpus Q, Corpus QBA, and Corpus QAA. All samples in the training data for each corpus were used for training without considering bagging and boosting.

Ensemble models based on classification results were designed to classify questions using the average values of the probability predicted by the same types of ANN models; for example, the classification results from multiple classifiers including the DNN, CNN, or LSTM models. The ensemble models based on classifications results were grouped into two types: (a) homogeneous ensemble models using classification results from the same type of ANN models (e.g., the average of two results from DNN models) and (b) heterogeneous ensemble models using classification results from different types of ANN models.

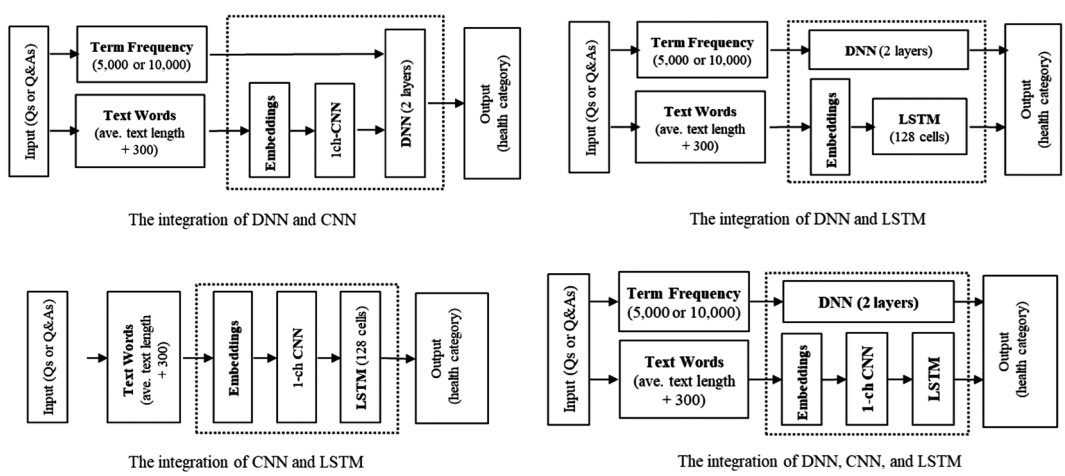
The accuracies of homogeneous ensemble models based on classification results were compared for DNN, single-channel CNN (1ch-CNN), four-channel CNN (4ch-CNN), and LSTM classifiers according to the number of classifiers (up to 20). Heterogeneous ensemble models based on classification results were designed as:

- DNN + CNN (1ch and 4ch)
- DNN + LSTM
- CNN (1ch and 4ch) + LSTM
- DNN + CNN (1ch and 4ch) + LSTM

Ensemble models based on the integrations of different ANN structures were designed by integrating different types of ANN structures for user-generated words as features (Figure 2). Only 1ch-CNN was used as a CNN model because no distinct difference was observed between one channel and four channels in terms of accuracy.

- DNN + 1ch-CNN: A DNN with two FCN layers featuring 128 and 64 nodes was integrated with a 1ch-CNN. The indexes for about 400 words were transformed into vectors with 100 dimensions for an embedding

Figure 2. Ensemble Models based on the Integration of Different ANN Structures



layer and used as the input to the CNN. The output of a flattening layer in CNN was merged with term frequency-based inputs, which were used as the input to the DNN.

- DNN + LSTM: DNN and LSTM with 128 cells were merged. The outputs from DNN and LSTM were merged for the input to the output layer.
- 1ch-CNN + LSTM: The output of the CNN pooling layer was used as input of LSTM with 128 cells.
- DNN + 1ch-CNN + LSTM: In the ensemble model including the aforementioned three types of structures.

4. Results

4.1 DNN, CNN, and LSTM models

In the DNN model, we employed frequency vectors as described in Table 2. The highest accuracy was produced when the classifier was trained with 10,000 terms that health consumers use. In addition, we observed that the larger the size of the features, the higher the accuracy. For example, the accuracy was 65% (.6502) for 100 features, 90% (.9006) for 2,000 features, 92% (.9150) for 5,000 features, and 94% (.9365) for 10,000 features. Classification with MeSH term features resulted in consistently lower accuracy than user-generated term features across all three corpus sets (Table 2).

Two CNN models (1ch and 4ch) and LSTM with 100-dimension word embeddings were compared. For 1ch-CNN, parameters were set as follows: kernel size of 4, 32 filters, and max-pooling size of 2. Two hidden layers with 128 and 64 nodes were implemented in the CNN models. The accuracies of these three models ranged from 89% to 92% across all three corpus sets. Of the two CNN models, the 1ch-CNN model with Corpus QBA produced the higher accuracy, 91% (.9120). However, we did not observe any distinct change in accuracy between 1ch-CNN and 4ch-CNN models. Some were higher in the 1ch-CNN model, whereas others were not. LSTM constantly showed better performance compared with CNNs across all corpus sets (Table 3).

4.2 Ensemble models

Two types of ensemble models were compared: ensemble models designed based on classification results and the integrations of different ANN structures for health consumer terms as features.

Table 2. Classification Accuracies for Corpora and Feature Type (Word Frequency)

Corpus	Features (*n*)	User-generated terms freq.	MeSh terms freq.
Corpus Q	5,000	.9086	.8641
	10,000	.9150	--
Corpus QBA	5,000	.9285	.8842
	10,000	**.9365**	--
Corpus QAA	5,000	.9237	.8748
	10,000	.9306	--

Table 3. Classification Accuracies of CNN and LSTM Models

Corpus	Features (*n*)	1 ch-CNN	4 ch-CNN	LSTM
Corpus Q	5,000	.8987	.8992	.9052
	10,000	.9010	.8961	.9062
Corpus QBA	5,000	.9062	.9043	.9140
	10,000	.9120	.9109	.9208
Corpus QAA	5,000	.8999	.9028	.9089
	10,000	.9105	.9084	.9169

Ensemble models based on classification results were further segmented into homogeneous or heterogeneous models to compare their accuracy. Accuracy of the ensemble models based on classification results may vary according to conditions. The lineup of the base classifiers can also affect the performance of a model (Bian & Wang, 2007). Thus, the two types of ensemble models based on classification results were further compared.

4.2.1 Ensemble models based on classification results (model averaging)

The averages of the classification probabilities from several ANN models were compared to decide final classification. Ensemble models classified each question into a category with the highest average probability by comparing the average values of two or three probabilities of belonging to each health category. For example, the average values of two probability values from DNN and CNN classifiers were compared for each classification in the DNN + CNN models, whereas three probability values were used in the ensemble model based on three classifiers, such as the DNN + CNN + LSTM model.

The ensemble models outperformed the individual models in accuracy for all corpora and features (Table 4). The ensemble models based on DNN and LSTM showed the highest accuracy in Corpus QBA with 10,000 features, which also outperformed the ensemble models based on the results from all three classifiers. Figure 3 shows the accuracy of the ensemble models based on classification results, comparing with the accuracy of individual classifiers. The results for Corpus QBA showed the best accuracy, followed by Corpus QAA and Corpus Q. Overall, 10,000 features were more effective in increasing accuracy than 5, 000 features.

4.2.2 Homogeneous and heterogeneous ensemble models based on classification results

Two type of ensemble models based on classification results—homogeneous and heterogeneous—were further evaluated and compared for Corpus Q with 5,000 features. For homogeneous ensemble models based on classification results, 20 classifiers per ANN model for Corpus Q with 5,000 terms were generated. The average accuracy of those classifiers was .9087 for DNN, .8960 for 1ch-CNN, .8968 for 4ch-CNN, and .9052 for LSTM. In general, the ensemble models showed higher accuracy compared to the average accuracy of the individual classifiers (Figure 4 and Appendix

Table 4. Accuracy of the Ensemble Classification Models Based on Classification Results

Corpus	Features (*n*)	DNN + 1ch-CNN	DNN + LSTM	1ch-CNN + LSTM	DNN + 1ch-CNN + LSTM
Corpus Q	5,000	.9197	.9206	.9106	.9225
	10,000	.9260	.9266	.9134	.9264
Corpus QBA	5,000	.9354	.9381	.9190	.9350
	10,000	.9424	***.9444***	.9263	.9412
Corpus QAA	5,000	.9269	.9297	.9128	.9271
	10,000	.9342	.9367	.9198	.9342

Figure 3. Accuracy Comparison of Individual Classifiers and Ensemble Models based on Classification Results

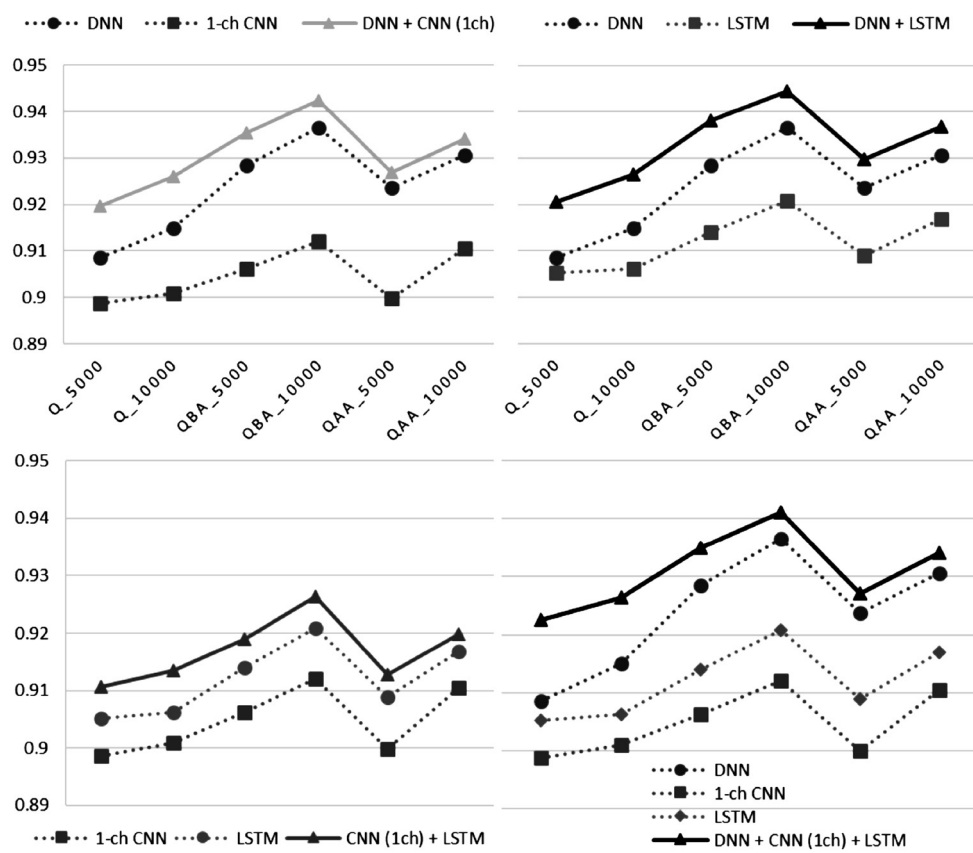
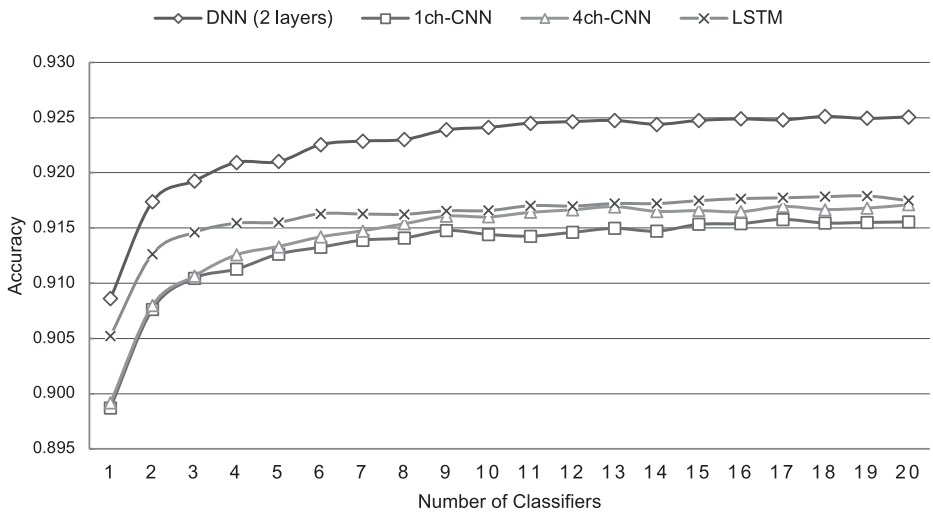


Figure 4. Accuracy Comparison of the Homogenous Ensemble Models Using Classification Results by the Number of Classifiers (Corpus Q with 5,000 Features)



A). Each individual model was randomly selected for the ensemble models. The ensemble models based on classification results tended to show higher accuracy as they used more classification results, although it was not guaranteed after nine classifiers (1ch-CNN), wherein it showed signs of convergence. All four ensemble models showed the largest improvements when two classification results were incorporated, and then the improvement rate in accuracy decreased. After using around five classifiers, the curves for accuracy improvement were likely to become flat.

Ensemble models based on DNN showed the highest accuracy and had a bigger gap relative to the other ANN models, followed by LSTM and CNN (4ch and 1ch). The 4ch-CNN model showed slightly higher accuracy than the 1ch-CNN model as the number of classifiers increased.

For heterogeneous ensemble models, classification results from more than two different types of ANN classifiers, which were combinations of DNN, CNN, and LSTM, were determined. In general, the number of classifiers was positively correlated with the accuracy of the ensemble models, similar to the homogenous ensemble models. The ensemble model constructed using two classifiers—DNN and LSTM—showed higher accuracy than the models based on CNN and LSTM. When more classification results were averaged, the classifications became more accurate. The ensemble models using classification results from DNN and 1ch-CNN classifiers were most effective regarding accuracy, whereas ensemble models using 1ch-CNN and LSTM were relatively less effective. The ensemble models including DNN classifiers showed relatively higher accuracy, whereas the ensemble models using CNN classifiers showed relatively lower accuracy. The accuracy of the heterogenous ensemble models based on two and three types of ANN classifiers is reported in Table 5 and Table 6,

Table 5. Accuracies of Heterogeneous Ensemble Models Using Classification Results of Two Different Types ANN Models by the Number of Classifiers (Corpus Q with 5,000 Features)

Num. of classifiers	DNN + 1ch-CNN	DNN + 4ch-CNN	DNN + LSTM	1ch-CNN + LSTM	4ch-CNN + LSTM
2	.9197	.9206	.9206	.9106	.9109
4	.9244	.9252	.9265	.9150	.9150
6	.9265	.9267	.9277	.9160	.9163
8	.9274	.9276	.9286	.9171	.9178
10	.9282	.9280	.9291	.9171	.9180
12	.9283	.9290	.9297	.9175	.9181
14	.9294	.9289	.9299	.9178	.9183
16	.9294	.9295	.9304	.9183	.9189
18	**.9296**	.9298	.9304	**.9185**	.9187
20	.9294	**.9299**	**.9304**	.9182	**.9192**

Table 6. Accuracies of Heterogeneous Ensemble Models Using Classification Results of Three Different Types ANN Models by the Number of Classifiers (Corpus Q with 5,000 Features)

Num. of classifiers	4ch-CNN + LSTM	DNN + 1ch-CNN + LSTM	DNN + 4ch-CNN + LSTM
3	.9133	.9225	.9220
6	.9163	.9248	.9245
9	.9182	.9264	.9269
12	.9181	.9270	.9277
15	.9186	.9275	.9278
18	**.9187**	**.9276**	**.9282**

respectively. Figure 5 represents the accuracy of the heterogeneous models.

When comparing the accuracy of homogenous and heterogeneous ensemble models based on classification results, the heterogeneous ensemble models showed relatively higher accuracy than the homogenous ensemble models in most cases. However, the accuracy of the heterogeneous ensemble models using the small numbers (e.g., 2, 4, 6) of classifiers based on CNN and LSTM was a little bit lower than that of the homogeneous models. The heterogeneous ensemble models using two ANN classifiers were compared with the homogenous ensemble models (Figure 6). All heterogeneous ensemble models using three different types of ANN classifiers showed higher

Figure 5. Comparisons of the Accuracy of Heterogeneous Ensemble Models Using Classification Results of ANN Models by the Number of Classifiers (Corpus Q with 5,000 Features)

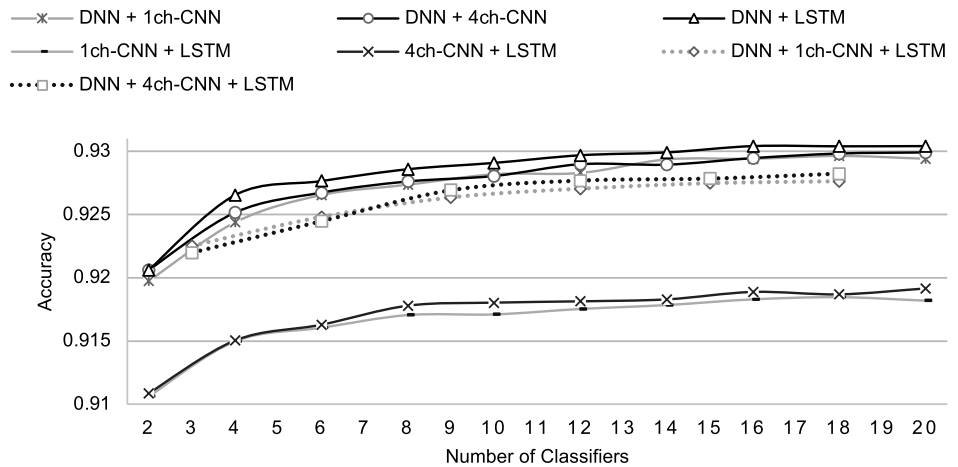
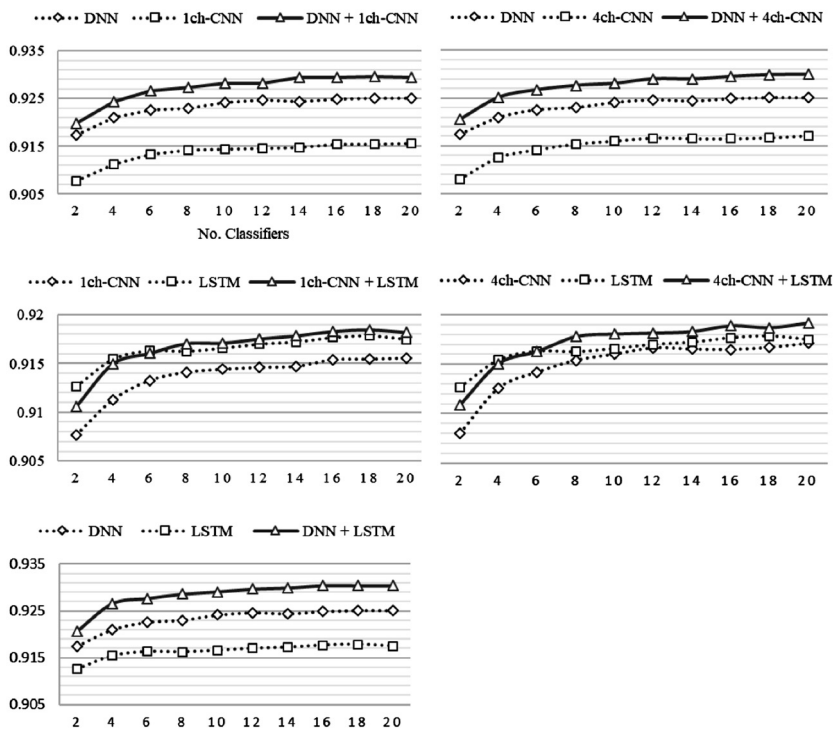


Figure 6. Comparisons of the Accuracy of Homogenous and Heterogeneous Ensemble Models Using Two ANN Models by the Number of Classifiers (Corpus Q with 5,000 Features)



accuracy than homogenous ensemble models (Figure 7).

4.2.3 Ensemble models based on the integration of different types of ANN structures

The accuracy of ensemble models using DNN, 1ch-CNN, and LSTM was compared (Table 7). We added word embeddings with 100 dimensions before CNN and LSTM layers. Using three feature sets and corpora, the DNN-based ensemble models demonstrated relatively higher accuracy than the CNN- or LSTM-based ensemble models. Corpus QBA with 10,000 features showed the highest accuracy of .9366 among the evaluated ensemble models.

The accuracies produced from all evaluated models exceeded 90% across the corpus sets. The models with 10,000 features showed relatively higher accuracy than those with 5,000 features. Accuracies exceeding 93% were observed in most cases with Corpus QBA and 10,000 features

Figure 7. Accuracy Comparison of Homogenous and Heterogeneous Ensemble Models Using Three ANN Models by the Number of Classifiers (Corpus Q with 5,000 Features)

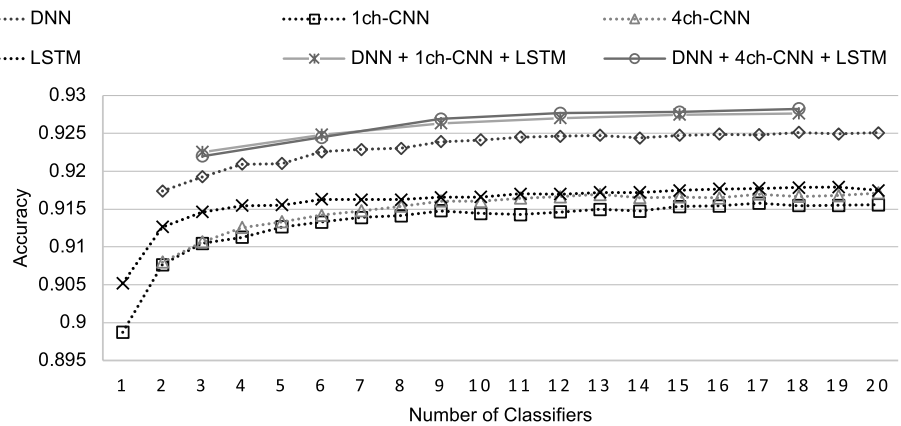


Table 7. Accuracies of the Ensemble Classification Models based on the Integration of ANN Heterogeneous Structures

Corpus	Features (*n*)	DNN + 1ch-CNN	DNN + LSTM	1ch-CNN + LSTM	DNN + 1ch-CNN + LSTM
Corpus Q	5,000	.9095	.9090	.9006	.9107
	10,000	.9176	.9154	.9033	.9175
Corpus QBA	5,000	.9242	.9299	.9117	.9311
	10,000	.9318	**.9366**	.9173	.9351
Corpus QAA	5,000	.9210	.9229	.9028	.9193
	10,000	.9288	.9309	.9121	.9289

except for the CNN + LSTM model. In particular, the ensemble model based on 1ch-CNN and LSTM produced the highest accuracy, 94% (.9366) in Corpus QBA.

In Figure 8, accuracy was compared between the individual classifiers and ensemble models based on the integrated structure of ANN models. Although the ensemble models did not exhibit lower accuracy than the individual ANN models, they also did not show synergy in the integration of structures. Accuracy was around the average of two classifiers or a little bit higher. When the DNN structure was embedded into ensemble models, the accuracy of the ensemble models showed a similar pattern to that of the DNN model.

The accuracy of two types of ensemble models, based on classification results or the integration of different ANN structures, was compared. The comparisons were based on the results presented in Tables 4 and 7. The accuracy of the ensemble models based on classification results outperformed the ensemble models based on the integration of heterogeneous ANN structures in all corpora, regardless of the number of features. Additionally, both types of ensemble models commonly produced the highest accuracy in corpus QBA (Figure 9).

Figure 8. Comparisons of the Accuracy of Individual Classifiers and Ensemble Models based on the Integrated Structure

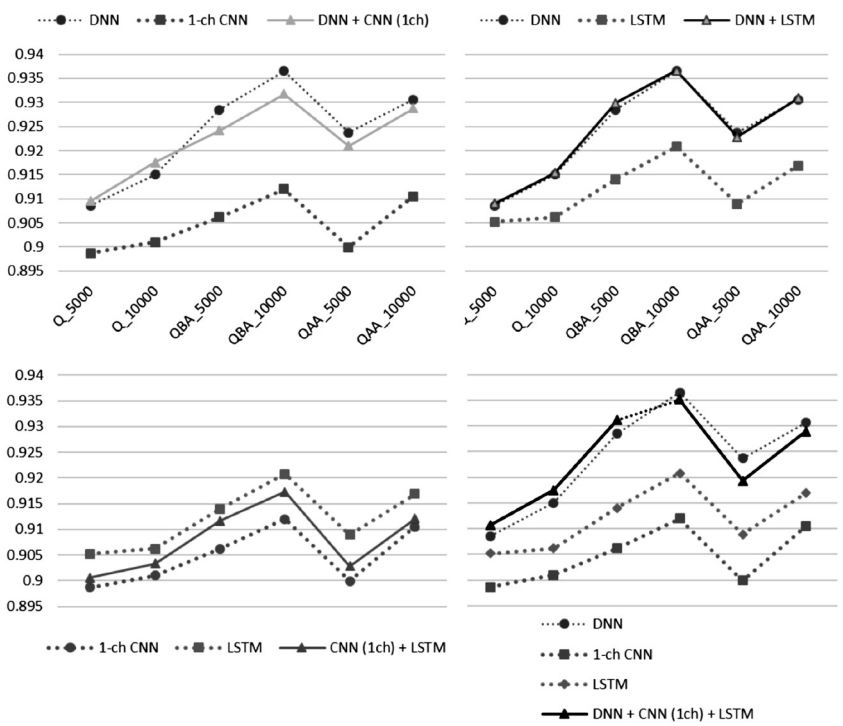
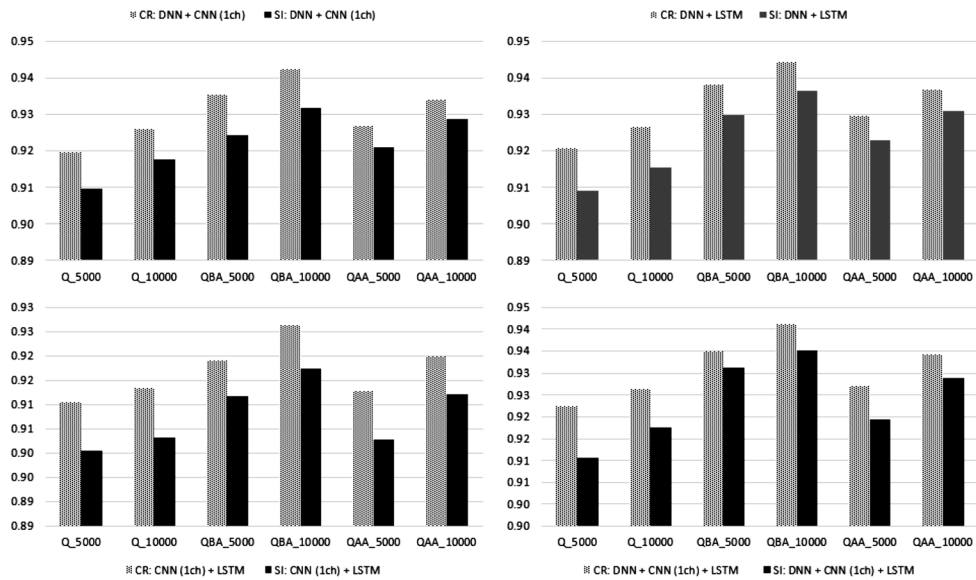


Figure 9. Accuracy Comparison of Ensemble Classification Models: Classification Results versus Structure Integration



5. Discussion and Conclusion

5.1 Discussion of findings

To identify the most optimal automated classification method that can reproduce health consumers' classification practices in social media, this study evaluated the classification accuracy of machine learning models, i.e., ANN and ensemble models. This study was motivated by not only the practical needs of automated organization of scaled online health resources on the web, but more importantly, to examine epistemology, which is a longstanding philosophical stance in IS. In the field of IS, KOSs such as classification systems have been developed or modified to meet users' needs from different communities (Greenberg, 2003; Pierre, 2001; Sarasohn-Kahn, 2008; Svenonius, 2000). This epistemic position has its strong roots in KOS research. Previous studies utilizing automated classification tasks have advanced mainly in the tech sector (McKinsey Global Institute, 2017) and were widely separated from this epistemological approach. On the other hand, IS scholars, particularly those interested in KOSs, have acknowledged the need to integrate an automatic approach to KOS issues due to ongoing transformations in knowledge production processes related to big data and Web 2.0. But few studies in IS have integrated automatic approaches to organizing domain knowledge. Our study evaluated advanced computational models from the standpoint of reinstating the value of user-centered approaches in classification of health resources. By evaluating the accuracy of computational models in reproducing health consumers' classification practices and utilizing their vocabularies as training data and features, our study examined the potential of machine

learning techniques from a bottom-up perspective rather than that of traditional health professionals. This study filled gaps in the existing literature by evaluating neural network models in classifying user-generated health information text uploaded to a social website.

The findings of this study indicate that features extracted from health consumers' lay language may correlate stronger with the social media corpus than a controlled vocabulary. Particularly, the text corpus of questions and best answers was most effective as a training set for ANN-based classifiers, showing the highest performance across all compared classification models including ensemble models. We speculate that the subject terms (e.g., depression, allergic reaction, and protein) in questions (e.g., "help! Allergic reaction to bug bites?") and best answers better served as a training dataset. In the context of social Q&As, questions are considered an embodiment of users' information needs in certain topics (Dervin, 1983; Shah et al., 2009) because they contain subject terms directly relevant to specific health issues or topics. So does the case of best answers because they are usually considered quality content with high relevance as evaluated by questioners (Gazan, 2011; Oh & Worrall, 2013). In contrast, we speculate that answers, which were not chosen as the best ones, tend to be less relevant to the question topics and include more random words, which may increase the entropy of the dataset. These results confirm that unfiltered user vocabularies could be good candidates for features when training an ANN model with social media data in consumer health domains. When comparing user-generated natural language and MeSH terms, the adequacy of user-generated terms as a training set for machine learning over established controlled vocabularies was further confirmed. Convergence of the prediction models were observed from user-generated terms, not from MeSH terms. This led us to conclude that user-generated terms would be more coherent in prediction models than a controlled vocabulary such as MeSH terms in reproducing classification practices in social web. Although MeSH terms were comparatively less effective, they still showed acceptable levels of performance, achieving accuracies ranging from 86% (.8641) to 88% (.8842) across different corpus sets. This implies that MeSH terms would be a good alternative in situations where it is not easy to identify user-generated terms, such as no available information about datasets. These results indicate that users' vocabulary could be more effective as features in classifying resources in social media although the results might be different depending on the applied field.

Regarding machine learning models, findings of this study confirm the reliable performance of ANN classifiers for user-generated health information in the context of social Q&A sites. The study examined multiple models of neural networks, including DNN, single-channel CNN, multichannel CNN, LSTM, and ensemble models consisting of multiple classifiers. All individual ANN models produced high-performance classification, showing accuracies of about 90%. But of those evaluated models, the ensemble models based on classification results outperformed other ensemble models based on the combination of different types of ANN structure models or individual classifiers.

Ensemble models based on classification results using heterogeneous classifiers were found to achieve higher accuracy than those using homogeneous classifiers. Ensemble models based on classification results, particularly using all three types of ANN models, showed improvements in accuracy. The number of classifiers also was positively correlated with classification performance. As the number of classifiers used increased, accuracy seemed to converge regardless of whether classifiers were homogeneous or heterogeneous. Between five and ten looked appropriate in terms of the number of classifiers. However, increases in the number of classifier types did not always improve accuracy. For example, the ensemble model based on two ANN models, DNN and LSTM, showed better accuracy than one based on all three types, DNN, CNN, and LSTM.

Model-averaging ensemble models used in this study can be compared with other types of ensemble models. The model-averaging ensemble model based on three heterogeneous ANN classifiers (DNN, CNN, and LSTM) was compared with a stacking ensemble model, which was trained with one hidden layer (128 nodes) based on inputs from three classifiers (Appendix B). The results were similar to the model-averaging ensemble model.

The ensemble models based on the integration of heterogeneous structures did not show explicitly better accuracy than the individual classifiers. We could not conclude that ensemble models based on the integration of heterogeneous structures are better than DNN classifiers in the context of user-generated health information classification. As for CNN models, we did not find any indicative difference in performance between one-channel and four-channel models. Although multiple channels were expected to generate better performance (Kim, 2014), we did not observe such effects in our experiment. We also found that LSTM did not result in higher accuracy than the frequency-based DNN model; however, LSTM was more effective than CNN in increasing accuracy.

5.2 Implications

This study made contributions to both research and practice. Regarding the literature, this study is one of a few empirical attempts to address automated classification methodology from an epistemological standpoint in IS. This attempt might draw attention from diverse research communities to KOS issues. Practically, this study contributed to finding the optimal approach to training neural networks for health consumer-generated resources in their own words. This study demonstrated how to construct adequate features from user language text and utilize them to identify optimized machine learning models for the classification of social media health information. The classifiers suggested in this study can assist in designing automated category systems in the health domain. The findings of this study could be useful for social Q&A sites or online health communities that utilize Q&A communication systems in suggesting or automatically categorizing users' questions or recategorizing existing posts in social media. The classifiers proposed herein can be also applicable in developing a deep-learning classification system for online forums, intelligent Q&A systems, and dialog records that assist online communication between patients and health practitioners.

The highest accuracy shown in this study, which was generated by the (DNN + LSTM) ensemble model based on classification results, is 94.44%. The classification model may not be an effective assistive tool in health-related fields where classification might be very critical and sensitive, but the model could be an assistant tool in classifying online health resources generated by health consumers.

5.3 Limitations

Different configurations in parameters could affect the performance of classification, such as the number of filters, filter sizes, and layers (Hughes et al., 2017). However, due to limited computing resources, this study did not include performance evaluations of different sets of parameters of dropout; the number of nodes, filters, and layers; the size of kernel or max-pooling; and different types of optimization methods. We also applied a limited number of words (5,000 and 10,000 for DNN) and a fixed length in a corpus (between 300 and 400 words per post for CNN and LSTM) as input features, which may have missed some words that would have affected the performance of CNN and LSTM. If the number of words for DNN and the length of input text for CNN and LSTM increased, LSTM might have produced a better performance. These limitations illustrate the need for further studies that involve more sophisticated tuning of different parameters using enhanced computing resources. Despite these limitations, this study represents a unique contribution to the application of ANNs with less-structured health resources that are generated by health consumers.

References

Abbas, J. (2010). *Structures for organizing knowledge: Exploring taxonomies, ontologies, and other schemas*. Neal-Schuman.

Agatonovic-Kustrin, S., & Beresford, R. (2000). Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. *Journal of Pharmaceutical & Biomedical Analysis, 22*(5), 717-727. https://doi.org/10.1016/S0731-7085(99)00272-1

Andersen, N., & Söderqvist, T. (2012). *Social media and public health research*. University of Copenhagen.

Apté, C., Damerau, F., & Weiss, S. (1994). Automated learning of decision rules for text categorization. *ACM Transactions on Information Systems, 12*(3), 233-251. https://doi.org/10.1145/183422.183423

Assefa, S. (2007). *Human concept cognition and semantic relations in the unified medical language system: A coherence analysis* [Doctoral dissertation, University of North Texas]. UNT Digital Library. https://digital.library.unt.edu/ark:/67531/metadc4008/m1/1/

Bian, S., & Wang, W. (2007). On diversity and accuracy of homogeneous and heterogeneous ensembles. *International Journal of Hybrid Intelligent Systems, 4*(2), 103-128. https://doi.org/10.3233/HIS-2007-4204

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*(2), 123-140. https://doi.org/10.1007/BF00058655

Brownlee, J. (2019). *Ensemble learning methods for deep learning neural networks*. Machine Learning Mastery. https://machinelearningmastery.com/ensemble-methods-for-deep-learning-neural-networks

Calefato, F., Lanubile, F., & Novielli, N. (2016). Moving to stack overflow: Best-answer prediction in legacy developer forums. In M. Genero (Chair), *Proceedings of the 10th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement* (pp. 1-10). Association for Computing Machinery. https://doi.org/10.1145/2961111.2962585

Cline, R. J. W., & Haynes, K. M. (2001). Consumer health information seeking on the Internet: The state of the art. *Health Education Research, 16*(6), 671-692. https://doi.org/10.1093/her/16.6.671

Dahlberg, I. (2006). Knowledge organization: A new science? *Knowledge Organization, 33*(1), 11-19.

Dervin, B. (1983, May). *An overview of sense-making research: Concepts, methods and results* [Paper presentation]. Annual Meeting of the International Communication Association, Dallas, TX, United States. http://communication.sbs.ohio-state.edu/sense-making/art/artdervin83.html

Dumais, S., Platt, J., Heckerman, D., & Sahami, M. (1998). Inductive learning algorithms and representations for text categorization. In K. Makki & L. Bouganim (Eds.), *Proceedings of the Seventh International Conference on Information and Knowledge Management* (pp. 148-155). Association for Computing Machinery. https://doi.org/10.1145/288627.288651

Efron, B., & Hastie, T. (2016). *Computer age statistical inference: Algorithms, evidence, and data science*. Cambridge University Press.

Er, O., Cetin, O., Bascil, S., & Temurtas, F. (2016). A comparative study on parkinson's disease diagnosis using neural networks and artificial immune system. *Journal of Medical Imaging & Health Informatics, 6*(1), 264-268. https://doi.org/10.1166/jmihi.2016.1606

Freund, Y., & Schapire, R. E. (1996). Experiments with a new boosting algorithm. In L. Saitta (Ed.), *ICML'96: Proceedings of the Thirteenth International Conference on International Conference on Machine Learning* (pp. 148-156). Morgan Kaufmann.

Gazan, R. (2011). Social Q&A. *Journal of the American Society for Information Science & Technology, 62*(12), 2301-2312. https://doi.org/10.1002/asi.21562

Golub, K. (2019). Automatic subject indexing of text. *Knowledge Organization, 46*(2), 104-121. https://doi.org/10.5771/0943-7444-2019-2-104

Golub, K., Soergel, D., Buchanan, G., Tudhope, D., Lykke, M., & Hiom, D. (2016). A framework for evaluating automatic indexing or classification in the context of retrieval. *Journal of the Association for Information Science & Technology, 67*(1), 3-16. https://doi.org/10.1002/asi.23600

Greenberg, J. (2003). Metadata and the world wide web. In M. J. Bates & M. N. Maack (Eds.), *Encyclopedia of Library & Information Science* (3rd ed., pp. 1876-1888). CRC Press.

Gross, T., & Taylor, A. G. (2005). What have we got to lose? The effect of controlled vocabulary on keyword searching results. *College & Research Libraries, 66*(3), 212-230. https://doi.org/10.5860/crl.66.3.212

Harper, F. M., Moy, D., & Konstan, J. A. (2009). Facts or friends? Distinguishing informational and conversational questions in social Q&A sites. In D. R. Olsen & R. B. Arthur (Chairs), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 759-768). Association for Computing Machinery. https://doi.org/10.1145/1518701.1518819

Hartmann, J., Huppertz, J., Schamp, C., & Heitmann, M. (2019). Comparing automated text classification methods. *International Journal of Research in Marketing, 36*(1), 20-38. https://doi.org/10.1016/j.ijresmar.2018.09.009

Hjørland, B. (2007). Semantics and knowledge organization. *Annual Review of Information Science & Technology, 41*(1), 367-405. https://doi.org/10.1002/aris.2007.1440410115

Hjørland, B. (2008). What is knowledge organization (KO)? *Knowledge Organization, 35*(2/3), 86-101. https://doi.org/10.5771/0943-7444-2008-2-3-86

Hjørland, B. (2014). Theories of knowledge organization—Theories of knowledge. *Knowledge Organization, 40*(3), 169-181. https://doi.org/10.5771/0943-7444-2013-3-169

Hjørland, B. (2018). Indexing: Concepts and theory. *Knowledge Organization, 45*(7), 609-639. https://doi.org/10.5771/0943-7444-2018-7-609

Hughes, M., Li, I., Kotoulas, S., & Suzumura, T. (2017). Medical text classification using convolutional neural networks. *Studies in Health Technology & Informatics, 235*, 246-250. https://doi.org/10.3233/978-1-61499-753-5-246

Ibekwe-Sanjuan, F., & Bowker, G. (2017). Implications of big data for knowledge organization. *Knowledge Organization, 44*(3), 187-198. https://doi.org/10.5771/0943-7444-2017-3-187

Jacob, P. (Ed.). (2014). *Text-based intelligent systems: Current research and practice in information extraction and retrieval.* Psychology Press.

Kalantari, A., Kamsin, A., Shamshirband, S., Gani, A., Alinejad-Rokny, H., & Chronopoulos, A. T. (2018). Computational intelligence approaches for classification of medical data: State-of-the-art, future challenges and research directions. *Neurocomputing, 276*(7), 2-22. https://doi.org/10.1016/j.neucom.2017.01.126

Kamel Boulos, M. N., & Wheeler, S. (2007). The emerging Web 2.0 social software: An enabling suite of sociable technologies in health and health care education. *Health Information & Libraries Journal, 24*(1), 2-23. https://doi.org/10.1111/j.1471-1842.2007.00701.x

Khan, J., Wei, J. S., Ringnér, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R., Peterson C., & Meltzer, P. S. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine, 7*(6), 673-679. https://doi.org/10.1038/89044

Kim, S. (2013). An exploratory study of user-centered indexing of published biomedical images. *Journal of the Medical Library Association, 101*(1), 73-76. https://doi.org/10.3163/1536-5050.101.1.011

Kim, Y. (2014). *Convolutional neural networks for sentence classification*. arXiv. https://arxiv.org/abs/1408.5882

Kim, S., Oh, J. S., & Oh, S. (2008). Best-answer selection criteria in a social Q&A site from the user-oriented relevance perspective. *Proceedings of the American Society for Information Science & Technology, 44*(1), 1-15. https://doi.org/10.1002/meet.1450440256

Kim, T.-Y., & Cho, S.-B. (2018). Web traffic anomaly detection using C-LSTM neural networks. *Expert Systems with Applications, 106*, 66-76. https://doi.org/10.1016/j.eswa.2018.04.004

Lewis, D. D., & Ringuette, M. (1994). A comparison of two learning algorithms for text categorization. In *Third annual symposium on document analysis & information retrieval* (pp. 81-93). University of Nevada.

Li, Q., & Lu, S. C. (2008). Collaborative tagging applications and approaches. *IEEE MultiMedia, 15*(3), 14-21. https://doi.org/10.1109/MMUL.2008.54

Lin, H., Jia, J., Guo, Q., Xue, Y., Li, Q., Huang, J., Cai, L., & Feng, L. (2014). User-level psychological stress detection from social media using deep neural network. In K. A. Hua, Y. Rui, & R. Steinmetz (Chairs), *Proceedings of the 22nd ACM international conference on multimedia* (pp. 507-516). Association for Computing Machinery. https://doi.org/10.1145/2647868.2654945

Liu, F., Antieau, L. D., & Yu, H. (2011). Toward automated consumer question answering: Automatically separating consumer questions from professional questions in the healthcare domain. *Journal of Biomedical Informatics, 44*(6), 1032-1038. https://doi.org/10.1016/j.jbi.2011.08.008

McCallum, A., & Nigam, K. (1998). A comparison of event models for naive Bayes text classification. In M. Sahami (Chair), *Learning for text categorization: Papers from the 1998 AAAI workshop (Technical Reports Vol. WS-98-05)* (pp. 41-48). Amer Assn for Artificial Press.

McKinsey Global Institute. (2017). *Artificial intelligence the next digital frontier*. https://www.calpers.ca.gov/docs/board-agendas/201801/full/day1/06-technology-background.pdf

McRoy, S., Jones, S., & Kurmally, A. (2016). Toward automated classification of consumers' cancer-related questions with a new taxonomy of expected answer types. *Health Informatics Journal, 22*(3), 523-535. https://doi.org/10.1177/1460458215571643

Messai, R., Simonet, M., Bricon-Souf, N., & Mousseau, M. (2010). Characterizing consumer health terminology in the breast cancer field. *Studies in Health Technology & Informatics, 160*(Pt. 2), 991-994.

Norton, M. (2010). *Introductory concepts in information science*. Information Today.

O'Reilly, T. (2007). What is Web 2.0: Design patterns and business models for the next generation of software. *Communications & Strategies, 1*(1), 17.

Oh, S., & Worrall, A. (2013). Health answer quality evaluation by librarians, nurses, and users in social Q&A. *Library & Information Science Research, 35*(4), 288-298. https://doi.org/10.1016/j.lisr.2013.04.007

Oh, S., Worrall, A., & Yi, Y. J. (2011). Quality evaluation of health answers in Yahoo! Answers: A comparison between experts and users. *Proceedings of the American Society for Information Science & Technology, 48*(1), 1-3. https://doi.org/10.1002/meet.2011.14504801269

Oh, S., Zhang, Y., & Park, M. S. (2016). Cancer information seeking in social question and answer services: Identifying health-related topics in cancer questions on Yahoo! Answers. *Information Research, 21*(3). http://www.informationr.net/ir/21-3/paper718.html

Peters, I. (2009). *Folksonomies. Indexing and retrieval in Web 2.0*. K. G. Saur. https://doi.org/10.1515/9783598441851

Pierre, J. M. (2001). *On the automated classification of web sites*. arXiv. https://arxiv.org/abs/cs/0102002

Poikonen, T., & Vakkari, P. (2009). Lay persons' and professionals' nutrition-related vocabularies and their matching to a general and a specific thesaurus. *Journal of Information Science, 35*(2), 232-243. https://doi.org/10.1177/0165551508098602

Sarasohn-Kahn, J. (2008). *The wisdom of patients: Health care meets online social media*. California Health Care Foundation. https://www.chcf.org/wp-content/uploads/2017/12/PDF-HealthCareSocialMedia.pdf

Sarker, A., & Gonzalez, G. (2015). Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *Journal of Biomedical Informatics, 53*, 196-207. https://doi.org/10.1016/j.jbi.2014.11.002

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys, 34*(1), 1-47. https://doi.org/10.1145/505282.505283

Seedorff, M., Peterson, K. J., Nelsen, L. A., Cocos, C., McCormick, J. B., Chute, C. G., & Pathak, J. (2013). Incorporating expert terminology and disease risk factors into consumer health vocabularies. *Pacific Symposium on Biocomputing,* 421-432.

Shah, C., Oh, S., & Oh, J. S. (2009). Research agenda for social Q&A. *Library & Information Science Research, 31*(4), 205-209. https://doi.org/10.1016/j.lisr.2009.07.006

Shiri, A. (2013). Linked data meets big data: A knowledge organization systems perspective. *Advances in Classification Research Online, 24*(1), 16-20. https://doi.org/10.7152/acro.v24i1.14672

SimilarWeb. (2019). *Answers.yahoo.com traffic overview*. https://www.similarweb.com/website/answers.yahoo.com#overview

Smiraglia, R. P. (2015). *Domain analysis for knowledge organization: Tools for ontology extraction*. Chandos.

Smiraglia, R. P., & Cai, X. (2017). Tracking the evolution of clustering, machine learning, automatic indexing and automatic classification in knowledge organization. *Knowledge Organization, 44*(3), 215-233. https://doi.org/10.5771/0943-7444-2017-3-215

Smiraglia, R. P., & Lee, H.-L. (Eds.). (2012). *Cultural frames of knowledge*. Ergon-Verlag.

Smith, C. A., & Wicks, P. J. (2008). PatientsLikeMe: Consumer health vocabulary as a folksonomy. In P. Dykes (Chair), *AMIA Annual Symposium proceedings* (pp. 682-686). American Medical Informatics Association.

Svenonius, E. (2000). *The intellectual foundation of information organization*. MIT press.

Tennis, J. T. (2008). Epistemology, theory, and methodology in knowledge organization: Toward a classification, metatheory, and research framework. *Knowledge Organization, 35*(2/3), 102-112. https://doi.org/10.5771/0943-7444-2008-2-3-102

U.S. National Library of Medicine. (2018). *Medical subject headings*. https://www.nlm.nih.gov/mesh/filelist.html

Weller, K. (2010). *Knowledge representation in the social semantic web*. Walter de Gruyter. https://doi.org/10.1515/9783598441585

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks, 5*(2), 241-259. https://doi.org/10.1016/S0893-6080(05)80023-1

Xu, W., & Rudnicky, A. (2000). Can artificial neural networks learn language models? In G. Dinghua (Chair), *Sixth International Conference on Spoken Language Processing* (pp. 202-205). International Speech Communication Association. https://www.isca-speech.org/archive/archive_papers/icslp_2000/i00_1202.pdf

Zeng, Q. T., & Tse, T. (2006). Exploring and developing consumer health vocabularies. *Journal of the American Medical Informatics Association, 13*(1), 24-29. https://doi.org/10.1197/jamia.M1761

Zhang, G. P. (2000). Neural networks for classification: A survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 30*(4), 451-462. https://doi.org/10.1109/5326.897072

Zhang, X., Wu, J., He, Z., Liu, X., & Su, Y. (2018). *Medical exam question answering with large- scale reading comprehension*. arXiv. https://arxiv.org/abs/1802.10279

Zhao, Y., & Zhang, J. (2017). Consumer health information seeking in social media: A literature review. *Health Information & Libraries Journal, 34*(4), 268-283. https://doi.org/10.1111/hir.12192

(Received: 2021/5/4; Accepted: 2021/7/29)

Appendix A

Accuracy of Homogenous Ensemble Models Using Classification Results by the Number of Classifiers (Corpus Q with 5,000 Features)

Num. of classifiers	DNN (2 layers)	1ch-CNN	4ch-CNN	LSTM
1	.9086	.8987	.8992	.9052
2	.9174	.9076	.9080	.9126
3	.9193	.9105	.9107	.9146
4	.9210	.9113	.9126	.9154
5	.9210	.9126	.9133	.9155
6	.9226	.9133	.9142	.9163
7	.9229	.9139	.9148	.9163
8	.9230	.9141	.9154	.9163
9	.9239	.9148	.9161	.9166
10	.9241	.9144	.9160	.9166
11	.9245	.9143	.9164	.9170
12	.9246	.9146	.9167	.9170
13	.9247	.9150	.9169	.9172
14	.9244	.9147	.9165	.9172
15	.9247	.9153	.9166	.9175
16	.9249	.9154	.9165	.9177
17	.9248	**.9158**	.9170	.9177
18	**.9251**	.9155	.9167	.9178
19	.9249	.9155	.9168	**.9179**
20	.9251	.9156	**.9171**	.9175

Note. The best accuracies are in bold.

Appendix B

Accuracy of Ensemble Models based on Classification Results (DNN + 1ch-CNN + LSTM)

Corpus	Num. of features (n)	Model averaging	Stacking
Corpus Q	5,000	.9225	.9143
	10,000	.9264	.9229
Corpus QBA	5,000	.9350	.9351
	10,000	.9412	.9413
Corpus QAA	5,000	.9271	.9282
	10,000	.9342	.9342

圖書資訊學刊　第20卷第1期　（2022.6）　頁1-30
https://doi.org/10.6182/jlis.202206_20(1).001

集成式人工神經網絡模型於分類實務之可行性：以社群媒體之健康消費者資訊分類為例

Practicability of Ensemble Artificial Neural Network Models for a Classification Task: An Optimal Approach for Reproducing Classification Practices of Health Consumers Generated Text on Social Media

Sukjin You[1], Min Sook Park[2], Soohyung Joo[3]

摘　要

本文運用人工神經網絡（Artificial Neural Network, ANN）模型，再現社群媒體中健康資訊分類實務之準確性。本研究透過Yahoo!Answers健康類別之問答，提取健康資訊術語，並輔以醫學主題詞表（MeSH terms），訓練並比較數種類型的ANN模型和集成式模型的效能。研究顯示，ANN模型分類準確率約90%；其中，深度神經網絡（Deep Neural Network, DNN）與卷積神經網絡（Convolutional Neural Network, CNN）和長短期記憶模型（long short-term memory, LSTM）相比，分類表現更佳。基於分類結果的集成模型不僅優於以基於異質ANN結構的集成模型，也優於單一深度學習模型；本研究也發現問題和最佳答案的組合是最有效的訓練集，並可以建構準確的預測模型。研究結果顯示，ANN模型可有效輔助分類健康消費者以自然語言生成之線上健康資訊。

關鍵字：自動分類、深度學習、人工神經網絡、集成分類模型、知識組織

[1,2] 美國威斯康辛大學密爾瓦基分校資訊學院
School of Information Studies, University of Wisconsin at Milwaukee, Wisconsin, USA
[3] 美國肯塔基大學資訊科學系
School of Information Science, University of Kentucky, Lexington, Kentucky, USA
* 通訊作者Corresponding Author: Min Sook Park, E-mail: minsook@uwm.edu

註：本中文摘要由圖書資訊學刊編輯提供。
以APA格式引用本文：You, S., Park, M. S., & Joo, S. (2022). Practicability of ensemble artificial neural network models for a classification task: An optimal approach for reproducing classification practices of health consumers generated text on social media. *Journal of Library and Information Studies, 20*(1), 1-30. https://doi.org/10.6182/jlis.202206_20(1).001
以Chicago格式引用本文：Sukjin You, Min Sook Park, and Soohyung Joo. “Practicability of ensemble artificial neural network models for a classification task: An optimal approach for reproducing classification practices of health consumers generated text on social media.” *Journal of Library and Information Studies* 20, no. 1 (2022): 1-30. https://doi.org/10.6182/jlis.202206_20(1).001