Document representation based on probabilistic word clustering in customer-voice classification

Younghoon Lee yhoon.lee@dm.snu.ac.kr
Seokmin Song seokmin@dm.snu.ac.kr
Sungzoon Cho zoon@dm.snu.ac.kr

August 8, 2017

Abstract

As is widely known, customer-voice data plays an important role in various fields such as marketing, product planning or quality assurance. However, a few problems are associated with the classification of customer-voice data mainly due to the manual processes involved. Thus, this study involved focusing on building automatic classifiers for customer-voice data with newly proposed document representation methods based on neural embedding and probabilistic word clustering approach. Semantically similar terms are clustered into a common cluster through clustering the words generated from neural embedding by considering the membership strength of each word with each cluster derived from a probabilistic clustering method such as fuzzy C-means or Gaussian mixture model. It is expected that the proposed method is suitable for the classification of customer-voice data that are composed of unstructured text by considering the membership strength. The results indicated that the proposed method achieved an accuracy of 89.24% with respect to representational effectiveness and 87.76% accuracy with respect to the classification performance of customer-voice data consisting of 12 classes. Additionally, the method provided an intuitive interpretation for the generated representation.

1 Introduction

Customer-voice (Voice of the customers, VOC) is a term that denotes the feelings of customers regarding their experience with a product, service, and/or business. Explicit complaints and requirements as well as the unsatisfied needs of customers and overall satisfaction are inherent in customer-voice. Thus, several companies attempt to identify and respond to customer needs and expectations with the customer-voice analysis [20, 41].

The customer-voice analysis provides important outputs and benefits for product or service developers. It provides a detailed understanding of the customer requirements and a common language for a team to proceed forward in the product development process. Additionally, it could be a key input for the setting of appropriate design specifications for a new product or service and a highly useful springboard for product innovation [15, 17].

As mentioned above, customer-voice plays an important role in various fields and is used in various departments. Hence, it is important to categorize customer-voice data and deliver it to relevant

departments and responsible individuals. For instance, the categorization of customer-voice data of mobile device into system / user interface / design and appearance categories allows it be delivered to proper departments and also provides overall information on customer-voice distribution by function. Therefore, it is necessary for customer-voice data to be classified into functional categories prior to analyzing the same. Figure 1 shows a summary of the customer-voice data analysis process.

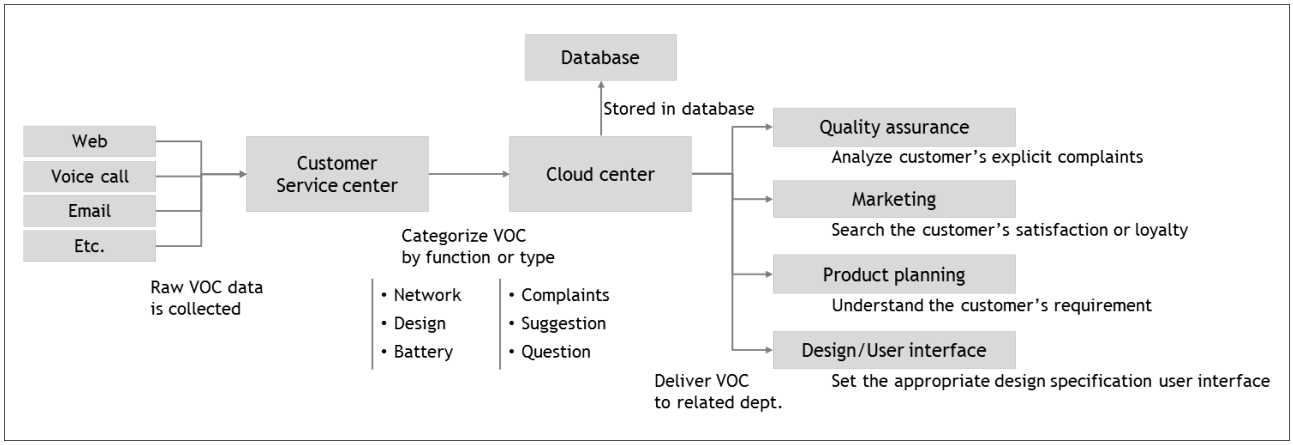


Figure 1: Summary of customer-voice data analysis process

In most companies, there are few problems with respect to the customer-voice classification mainly because the classification is performed manually. The first problem relates to the lack of consistency. Classification tasks are performed by several individuals, and thus, the results vary with the individuals and could result in the necessity for additional steps to correct the inconsistency. The second problem relates to the time-consuming nature of the classification. It is sometimes an urgent issue to respond to customer-voice especially when a quality assurance problem is involved. The time consumed on the classification task causes a delay in analyzing customer-voice and requires an immediate response. The last problem relates to resource management. Unnecessarily devoting human resources into a classification task can result in a lost opportunity to allocate these human resources into more important tasks and to optimize human resource management. Thus, this study focused on building an automatic classifier for customer-voice data with newly proposed methods categorizing step in figure 1. The aforementioned problems were effectively addressed by utilizing an appropriate classifier. The consistency of classification results and the removal of unnecessary time consumed are helpful in human resource management.

The customer-voice data in the study was gleaned from across a variety of channels including phone, e-mail, or web, and it was stored in a text document as shown in table 1. Thus, the classification of customer-voice data is considered as a subset of a document classification problem. Representation of a document is an essential part of a task in document classification. This study focused on suggesting an advanced document representation method that is appropriate for customer-voice data. Document representation for customer-voice data requires a better performance than previous methods as well as providing representational interpretability since it was analyzed for various purposes after the classification task.

The most common document representation method involves a bag-of-words or bag-of-n-grams in

Table 1: Example of customer-voice data

Customer-voice example	Type	Function
there are other problems, the most of which involves display brightness... Indoor or outdoors, I couldn't see the screen very well, especially under the sunlight.. i was anticipating LG to recognize this problem and plan to software updates. I am disappointed. I always turn on the auto brightness mode. Then when does day-lighting mode work?	Complaint	Display
Hi i am currently using G5. When I took pictures, they were saved in the sd card until now. Then suddenly it said that it couldn't save the photo and was unable to move photos either. I don't think it is the problem of the sd card storage space or file recognition problems. I removed and inserted sd card few times, but it didn't work.. I look forward to your response. Thanks.	Question	Camera and Gallery
I got g5 and have a question about camplus. After mounting camplus, i press shutter button, then the backup battery is activated and the main battery begins to be charged. First question is whether of not I can deactivate backup battery function of camplus? Second question is that.... Even when the main battery is completely charged, camplus backup battery (?) is still active. Then camplus battery is still used even when the main battery is full?	Question	Battery and Charging
Hello I am currently using G5. I got various accessories as well as G5. Then the VR device leave much to be desired. I expected you to support sw update soon.. but there were no updates yet. That is the reason why I wrote this email.. when I wear the VR device.. the view display is rotated as my head is moved. when I want to pin the front view i tilt my head. however, I want to watch movie lying down, then it always shows only the ceiling ..... could you support s/w update to solve this problem?	Suggestion	Accessory

which the document is fundamentally represented by the counts of word occurrences within a document due to its simplicity and efficiency [1, 28, 38]. However, these methods involve many limitations since they suffer from the sparsity of vectors and high dimensionality. Hence, various dimension reduction methods and feature extraction methods are described in section 2 to solve the sparsity and dimensional problem. However, these methods may lose the innate interpretation involved in the bag-of-words approach and continue to ignore the contextual information in the words in the document. In order to overcome the fore-mentioned limitations, the document-to-vector(doc2vec) model [26], one of the neural embedding approaches, is considered, which is an extension of the word-to-vector(word2vec) model [30, 31] in which words are presented in a continuous vector space considering the contextual information of words in a document. Irrespective of these advantages, this approach could not provide an intuitive interpretation of the document representation since the document was also represented in a continuous vector space and generated from a neural network model.

Furthermore, the bag-of-concepts approach leads to a better representational performance and intuitive interpretation when compared with those of previous studies [21]. With respect to the bag-of-concepts approach, semantically similar terms are clustered into a common concept by clustering the word generated from neural embedding. Similarly, Paniagua(2015) also utilize word cluster based on word generated from neural embedding for named entity recognition task. Those studies are almost

starting studies of applying word clustering approach based on vec architecture [40].

Another word clustering approach based on their similarity is used in many studies of text mining. Ghayoomi (2012) and Sagae (2009) construct parser based on word clustering [16, 35]. And Saha(2008) utilize the word clustering for named entity recognition [36]. Mitrofanova(2009) and Bekkerman(2003) utilizes the word clustering in text categorization [32, 2].

However, previous approaches including recent studies based on neural embedding are based on the hard clustering method. Thus, it does not reflect the membership strength of each word with respect to each cluster. Based on observations in the study, words placed closest to the centroid of each cluster constitute meaningful keywords while words placed far from the centroid do not constitute meaningful keywords. Therefore, it is reasonable to differentiate between those words in constructing the word clustering. Therefore, in this study, an advanced representation method utilizing the neural embedding architecture and considering membership strength was proposed. The method was based on the probabilistic clustering method considering membership strength of each word with each cluster. The customer-voice data is composed of extremely unstructured text even including typo's. Hence, it was expected that the proposed method is robust with respect to unstructured customer-voice data by considering the membership strength of those words and exhibits a better representation performance when compared with that of previous methods. The proposed method was applied in the representation and classification of customer-voice data of mobile devices collected from LG Electronics during the past three years to verify the performance and adaptability of the proposed method.

The rest of this paper is structured as follows. Section 2 discusses various studies of the document representation method including the word clustering based method that was improved upon in this study. Additionally, the proposed improved document representation method is presented in Section 3. Section 4 presents the data description and experimental results for representational effectiveness and classification performance of the proposed method. Section 5 provides the conclusions and the discussion and directions for future work.

2 Related work

As described above, document representation method is a key step in the document classification problem. This section reviews the major document representation methods. Many text and sentiment classifiers are still solely based on different sets of words contained in documents such as the bag-of-words or bag-of-n-grams without considering sentence and discourse structure or meaning. It is a straightforward method and provides an intuitive interpretation. However, these approaches are limited when a large number of documents are involved. It could be composed of huge size of dimension and too spare to measure the proximity between documents [23, 45].

Latent semantic analysis (LSA) [12], probabilistic latent semantic analysis (pLSA) [5] and more comprehensive method based latent Dirichlet allocation (LDA) were suggested [3] to reduce the dimension and select more discriminative feature. However, these techniques could lose the innate interpretability of bag-of-concepts approach since each of the features is considered as an artificial concept. Additionally, the method suffers from few disadvantages since it continues to be based on

word co-occurrences. It ignores the semantic relevance among words and does not consider the context information to a lesser extent when compared with the bag-of-words method. Furthermore, the inference process is too sensitive to the initial condition and especially with respect to the LDA based model.

Additionally, word2vec, one of the neural embedding approaches, is based on the assumption of the distributed hypothesis, which implies that words occurring in a similar context tend to have similar meanings [18]. Based on this assumption, word2vec uses a neural network model such as skip-gram or continuous bag of word (CBOW) that predicts the neighboring words of input words [26, 30]. In word2vec, a neural network model is first trained with respect to the optimization function $\frac{1}{T}\sum_{t=k}^{T-k} log(p(\omega_t|\omega_{t-k},...,\omega_{t+k}))$ in CBOW or $\frac{1}{T}\sum_{t=k}^{T-k} log(p(\omega_{t-k},...,\omega_{t+k}|\omega_t))$ in skip-gram when T denotes the number of words, and k denotes the window size of neighboring words. Hidden nodes could then be used as the representations of words w_t. The most important aspect of word2vec is that words with similar meaning are located close to each other in vector space. The word2vec model can be utilized to construct dense document vectors with reasonable dimensions when compared to the bag-of-words approach in which the dimension and sparsity of a document vector can increase significantly.

Various document representation methods are suggested based on the word2vec model. Even a simple representation method in which average word vectors are contained in document shows a good representation performance [46]. A promising representation method based on word2vec model corresponds to the doc2vec model. The doc2vec model utilizes contextual information of words and documents to represent a document d_j as well as w_i in a vector space by training the optimization function $\sum_{i=1}^{V} \log p(w_i|w_{context}) + \sum_{j=1}^{D}\sum_{i=1}^{V} \log(w_i|d_j)$ when V denotes the number of words, and D denotes the number of documents. With respect to contextual information, the document vectors with similar contextual information are placed close to each other in a vector space. The representational performance of doc2vec model exhibits a better performance than those of the bag-of-words approaches and word2vec averaging method [9]. However, the dimension of the vector generated from doc2vec does not provide any interpretation since each document vector is trained through a neural network, and each value of the vector represents only the strength of the connection between the nodes. Consequently, it is difficult to comprehend what each feature of a document vector represents in terms of the document contents.

The bag-of-concepts approach, one of the word clustering based documentation representation method, combines the advantages of previous studies. Semantically similar terms are clustered into a common concept through clustering the words generated from neural embedding architecture, thereby incorporating the impact of semantically similar words on preserving document proximity. Document vectors are subsequently represented by the frequencies of these concepts. According to the experimental results, the bag-of-concepts approach provides better representational performance when compared to those of previous studies including bag-of-words, doc2vec, and LSA while simultaneously proving representation interpretability [21].

Similarly, Paniagua(2015) utilize word vectors and the word clusters generated by the neural

embedding architecture to add the best of both in the feature set of documents [40]. And, in Mitrofanova(2009), a set of key words describing major topics of the plot was assigned to each text, clusters of words with similar distributions were created for each key word based on word vector model which are utilizing co-occurrence matrix [32, 37]. Moreover, Saha(2008) constructs word clustering based cosine similarity for named entity recognition task [36]. Lastly, Bekkerman(2003) more directly compared the simple bag-of-words approach and the word cluster based document representation approach, then prove the effectiveness of word clustering based document representation.

Those studies, however, utilize hard clustering method such as K-means, K-medoids or spherical K-means clustering not to consider membership strength of each word with each cluster. Therefore, in the present study, an advanced document representation method utilizing neural embedding architecture based on probabilistic clustering method was proposed to capture the membership strength of each word. The utilized probabilistic clustering methods included fuzzy C-means (FCM) clustering method [19] and Gaussian mixture model (GMM) clustering method [13].

The fuzzy C-means algorithm attempts to partition a finite collection of n elements $X = \{X_1, ..., X_n\}$ into a collection of c fuzzy clusters with respect to a specified criterion. Given a finite set of data, the algorithm returns a list of c cluster centers $C = \{C_1, ..., C_c\}$ and a partition matrix $W = w_{ij} \in [0, 1]$, $i = 1, ..., n$, $j = 1, ..., n$ where each element w_{ij} specified the degree to which element X_i belongs to cluster C_j. The fuzzy C-means algorithm aims to minimize an objective function as follows:

$$\underset{C}{\operatorname{argmin}} \sum_{i=1}^{n} \sum_{j=1}^{c} w_{ij}^{m} dist^2(x_i, c_j)$$

where

$$w_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{dist(x_i, c_j)}{dist(x_i, c_k)} \right)^{\frac{2}{m-1}}}$$

In this study, the cosine distance was used as a distance measure in a manner similar to the previous approaches.

A Gaussian mixture model is a parametric probability density function that is represented as a weighted sum of Gaussian component densities. In a multivariate distribution, $p(x|\theta)$ is defined as a finite mixture model with J components, and each component is a multivariate Gaussian density defined with parameters $\theta_j = \{\mu_j, \Sigma_j\}$ as follows:

$$p(x|\theta) = \sum_{j=1}^{J} \alpha_j p_j(x|z_j, \theta_j),$$

$$p_j(x|\theta_j) = \frac{1}{(2\pi)^{d/2} |\Sigma_j|^{1/2}} e^{-\frac{1}{2}(x-\mu_j)^t \Sigma_j^{-1} (x-\mu_j)}$$

and $\alpha_j = p(z_j)$ denote the mixture weight representing the probability that a randomly selected x was generated by components J, and $\Sigma_{j=1}^{J} \alpha_j = 1$. After each parameter was calculated using the Expectation-Maximization (EM) algorithm, the membership weight of data point is computed as follows:

$$w_{ij} = p(z_{ij} = 1|x_i, \theta) = \frac{p_j(x_i|z_j, \theta_j) \cdot \alpha_j}{\sum_{m=1}^{J} p_m(x_i|z_m, \theta_m) \cdot \alpha_m}$$

In this study, w_{ij} is considered and is computed in fuzzy C-means clustering method and Gaussian mixture model clustering method as the membership strength of each word with each cluster.

3 Proposed method

As we mentioned above, previous word clustering based approaches has a limitation related to reflecting the membership strength of words with each cluster. That is, previous approaches represents a document based on the hard clustering method and does not differentiate in terms of frequency count as to whether a word is located closest to the centroid of each cluster or located far from the centroid.

Words are clustered in customer-voice data of mobile device collected from LG Electronics by spherical K-means method [47] in a manner identical to that in a previous study to show the limitation of not considering the membership strength of a word with a cluster. With respect to the spherical K-means method, data located near each centroid is considered to exhibit a strong membership with each centroid. Table 2 & Table 3 show the lists of words in the 7th cluster among 70 clusters with respect to cosine distance from the centroid. A close look at the 7th cluster indicates that it may contain words related to "water damage or breakage." Additionally, it is also revealed that words located near centroid, such as "rust," "humidity," and "LCD," are meaningful keywords to clearly represent the property of a cluster while other words located far from the centroid, such as "think," "daily," and "terminal," appear as relatively general words that are not strongly related to the "water damage or breakage" topic. A domain expert of LG electronics was involved in the study and shared the same opinion as the study observations.

Table 2: Word list located closest to centroid

Word	Distance	Word	Distance	Word	Distance
rust	0.7540	*careful*	0.7935	*broken*	0.8133
mistake	0.7818	*carelessness*	0.8099	*part*	0.8137
humidity	0.7851	*LCD*	0.8110	*dent*	0.8144
throw	0.7892	*tempered glass*	0.8123	*appearance*	0.8189

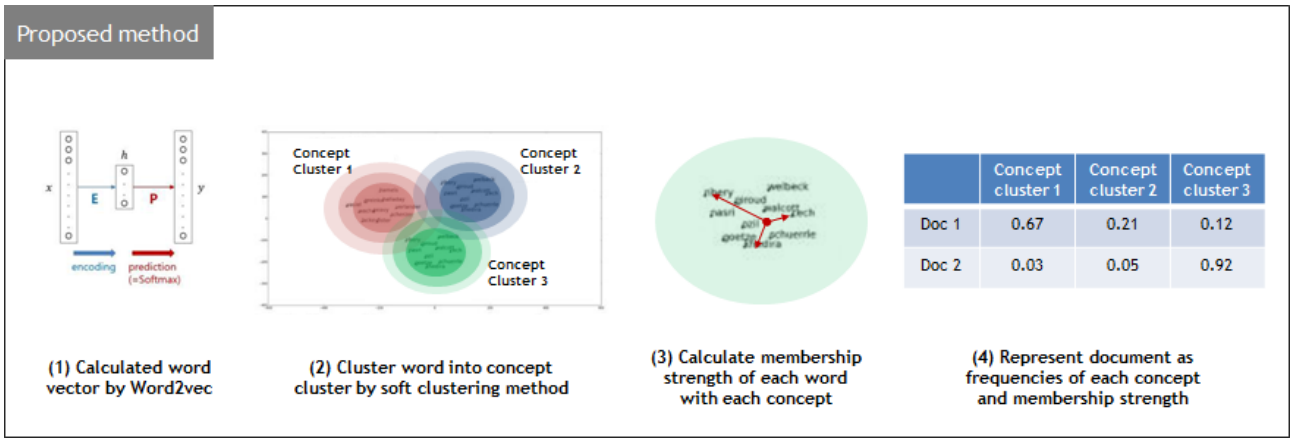
Table 3: Word list located far from centroid

Word	Distance	Word	Distance	Word	Distance
integrated	1.1593	*sticker*	1.1260	*do*	1.1132
just	1.1479	*two*	1.1206	*ambiguous*	1.1017
pay	1.1295	*tear*	1.1192	*grudge*	1.0930
terminal	1.1286	*daily*	1.1180	*think*	1.0912

Hence, it is reasonable to differentiate between words in the frequency count. That is, words exhibiting a strong membership with clusters need to make a higher impact in the frequency count

since these words better represent the property of the cluster. The consideration if membership strength is expected to increase the impact of meaningful keywords in document representation, and it is expected that the proposed representation method will be more robust with respect to noisy words since noisy words can have lower impact in the frequency count. In this study, two soft clustering methods, namely the fuzzy C-means clustering and Gaussian mixture model, are applied to measure the membership strength of each word with clusters. The application of the soft clustering method enabled the measurement of the member strength of words by w_{ij} in these methods. In the fuzzy C-means clustering method, w_{ij} denotes the degree to which $word_i$ belongs to cluster C_j, and w_{ij} denotes the probability that the $word_i$ is generated from the distribution of cluster C_j in the Gaussian mixture model method.

The following figure 2 summarizes the proposed document representation method. The proposed document representation method is calculated as follows:



	Concept cluster 1	Concept cluster 2	Concept cluster 3
Doc 1	0.67	0.21	0.12
Doc 2	0.03	0.05	0.92

Figure 2: Document representation based on probabilistic word clustering

Formally, let the set of documents $D = \{d_1, ..., d_N\}$ where N denotes the number of documents. The set of words $W = \{w_1, ..., w_n\}$ where n denotes the total number of words in D. Additionally, c denotes the number of clusters user define, and cf_{ijk} denotes the frequency of $word_i$ that is included in $cluster_j$ in $document_k$. Additionally, df_j denotes the number of documents containing words included in $cluster_j$, and $dist(a, b)$ denotes the cosine distance between a and b. Furthermore, $Centroid_j$ denotes the centroid of $cluster_j$.

Definition 1 (Word vector). *Word vector w_i is h-dimensional vector that represents each word, and h denotes the number of hidden nodes user define in the neural embedding model.*

Definition 2 (Membership strength). *Membership strength m_{ij} denotes the scalar value that represents the membership strength of $word_i$ with $cluster_j$ in which $m_{ij} \in [0, 1]$.*

Definition 3 (Document vector prior to normalization). *Document vector before normalization V_k denotes the j-dimensional vector of document, and formally corresponds to $V_k = [v_{k1}, ..., v_{kj}, ..., v_{kc}]$*

where $j = 1, ..., c$, $k = 1, ..., N$.

Definition 4 (Document vector). *Document vector* DV_k *corresponds to the j-dimensional vector of document with normalization that represents the final document vector, and formally corresponds to* $DV_k = [dv_{k1}, ..., dv_{kj}, ..., dv_{kc}]$ *where* $j = 1, ..., c$, $k = 1, ..., N$.

Step 1. *Calculate vector dimension of each word* w_i *using the neural embedding model.* w_i *is calculated by optimizing the function* $\sum_{i=1}^{V} \log p(w_i|w_{context})$.

Step 2. *Clustering all words* w_i *and Calculate membership strength* m_{ij} *for i,j with fuzzy C-means and Gaussian mixture model.*

With fuzzy C-means clustering method :

Apply fuzzy C-means method to calculate the membership strength m_{ij} *by the equation as follows:*

$$m_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{dist(w_i, Centroid_j)}{dist(w_i, Centroid_k)} \right)^2} \tag{1}$$

where $i = 1, ..., n$, $j = 1, ..., c$, *while minimizing an objective function as follows:*

$$\underset{C}{\operatorname{argmin}} \sum_{i=1}^{n} \sum_{j=1}^{c} w_{ij}^{m} dist^2(x_i, c_j)$$

With Gaussian mixture model as follows:

Apply the Gaussian mixture model to calculate the membership strength m_{ij} *by the equation as follows:*

$$m_{ij} = \frac{p_j(x_i|z_j, \theta_j) \cdot \alpha_j}{\sum_{k=1}^{J} p_k(x_i|z_k, \theta_k) \cdot \alpha_k} \tag{2}$$

where $i = 1, ..., n$, $j = 1, ..., c$, *while* $p(x|\theta)$ *is defined as a finite mixture model with J components, and each component is a multivariate Gaussian density defined with parameter* $\theta_j = \{\mu_j, \Sigma_j\}$ *as follows*

$$p(x|\theta) = \sum_{j=1}^{J} \alpha_j p_j(x|z_j, \theta_j),$$

$$p_j(x|\theta_j) = \frac{1}{(2\pi)^{d/2} |\Sigma_j|^{1/2}} e^{-\frac{1}{2}(x-\mu_j)^t \Sigma_j^{-1} (x-\mu_j)}$$

and $\alpha_j = p(z_j)$ *denote the mixture weight that represents the probability that a randomly selected x is generated by components J, where* $\Sigma_{j=1}^{J} \alpha_j = 1$. *Each parameter is updated by the EM algorithm.*

Step 3. *Calculate the document vector prior to normalization* $V_k = [v_{k1}, ..., v_{kj}, ..., v_{kc}]$ *by the*

following equation:

$$v_{kj} = \sum_{i} (cf_{ijk} \times m_{ij}) \tag{3}$$

where $i = 1, ..., n,\ j = 1, ..., c,\ k = 1, ..., N.$

Step 4. *Calculate the document vector* $DV_k = [dv_{k1}, ..., dv_{kj}, ..., dv_{kc}]$ *by the following equation:*

$$dv_{kj} = \frac{V_{kj}}{\sum_j V_{kj}} \times \log \frac{N}{df_j} \tag{4}$$

where $j = 1, ..., c,\ k = 1, ..., N.$

In Step 1, the word vector w_i is calculated by the neural embedding model. As described previously, the number of dimensions of w_i corresponds to the number of hidden nodes of neural embedding model that is defined by the user. The membership strength m_{ij} for all i, j is calculated in Step 2. Two soft clustering methods, namely the fuzzy C-Means clustering method and Gaussian mixture model, are used. Equation (1) is used to calculate m_{ij} with the fuzzy C-Means method and Equation (2) is used to calculate m_{ij} with the Gaussian mixture model. In step 3, document vector prior to normalization is calculated by multiplying membership strength m_{ij} and cf_{ijk}, which is the frequency of $word_i$ that is included in the j_{th} cluster in the k_{th} document based on equation (3).

In step 4, each dimension is first divided by summing the entire dimension for normalizing the effect based on equation (4). The normalization effect is applied to create a robust document representation on the length of the document. As mentioned above, the customer-voice data is an extremely unstructured text with various lengths and the longer customer-voice data often contains a large amount of repetition. Without normalization, there is a problem that the customer-voice data of different lengths containing similar contents can be differently represented and classified into different categories. Second, $\log \frac{N}{df_j}$ is multiplied with each dimension according to equation (4) for Concept Frequency-Inverse Document Frequency (CF-IDF) effect used on the previously specified bag-of-concepts approach. According to the previous bag-of-concepts study, the CF-IDF corresponds to the weighting scheme that readjusts the count of concept based on its frequency in the entire corpus. If a certain concept occurs in every document in the corpus, it is considered as relatively unimportant, thus reducing its frequency [21].

4 Experiment

4.1 Data description

In this study, in order to verify the representational effectiveness and classification performance of our proposed method and its applicability on customer-voice data, customer-voice data of mobile devices collected from Mobile Communication (MC) department in LG Electronics is used. The data was collected between April 23, 2014 and March 23, 2017. The data were manually labeled by domain

experts in LG Electronics into 12 classes. In order to avoid a class imbalance problem, a similar number of customer-voice were collected from each class as shown in the table 4 below.

Table 4: Customer-voice data set collected from LG Electronics

Class	Number of Customer-voice data	Class	Number of Customer-voice data
OS upgrade	900	Network connection	900
Multimedia	900	Call & Message	900
Hard key & input error	900	Heating & Processing	900
Water-proof / Dust-proof	900	Battery & Power	900
Accessory	793	Appearance & Display	900
Security & Backup	900	User Interface	900
		Total 12 classes	**10,693**

4.2 Experiment setup

Two experiments are performed to verify the representational performance of our proposed method and its applicability on classification of customer-voice data. The first experiment is performed to analyze the representational effectiveness, and the second experiment is performed to measure the classification performance. In the experiments, representational effectiveness and classification performance based on the proposed document representation method are compared to those generated from the bag-of-words, word2Vec averaging, doc2vec, Topic vector, bag-of-concepts and Latent Semantic Analysis(LSA) approaches. The topic vector is an inferred topic proportion that is typically used as a topic feature to represent the document [6]. And LSA is the technique applying singular value decomposition(SVD) in term-frequency matrix to reduce the number of rows while preserving the similarity structure among columns [24]. The first experiment is similar to those in studies by Dai et al. [9]. In the fore-mentioned studies, triplets of documents were constructed in which two documents were chosen from the same class while the other document was selected from a different class. The experimental result was considered correct if the document calculated as most distant was from a different class. The dataset of the present study contains 12 different classes, and thus 132 unique combinations of the triplets are constructed. Additionally, 1000 triplets are created for each combination, and thus, the experiment is performed on 132000 triplets. The document classification task is performed in the second experiment. The classification result is considered as correct if the document is predicted as its actual class by the prediction model. A major voting ensemble model that is used in several studies [34, 4] is constructed for the classification task. The k-nearest neighbor [33, 7, 8], support vector machine [39, 43, 42], logistic regression [44], Gaussian Naive Bayes classifier [25, 10, 27], and neural network [29, 14] approaches were combined for the ensemble model. The proposed method, doc2vec method, and bag-of-concepts method were designed to share same window size of 8 and the number of hidden layers in training word vector (300) to minimize the impact of hyper-parameters in the overall experiments. Furthermore, the proposed method and bag-of-concepts method are influenced by the

number of clusters. Hence, in order to observe the impact on the overall experiments, several values were experimented with the number of clusters beginning from 20 until 200 with increments of 10.

4.3 Experiment results

4.3.1 Representational effectiveness

Table 5 and Figure 3 show the results of the representational effectiveness of various methods with respect to varying dimensions. As shown in figure3, the probabilistic word clustering based approach with the FCM method outperforms all other document representation methods in all dimensions. The reason for higher performance of the proposed method is that it captures the membership strength of a word with each cluster. The experiment result shows a higher absolute performance level given that the number of classes corresponded to 12. Additionally, the probabilistic word clustering based approach with the GMM method exhibits a lower performance than that of the probabilistic word clustering based approach with FCM approach. A reason for these results is the problem related to GMM assumptions in which the GMM method assumes that each data set is generated from each Gaussian distribution. However, it appears that the words used in the experiment did not fit the assumption of GMM very well. However, in spite of these problems, the probabilistic word clustering based approach with GMM method exhibits a performance also outperforms all other representation methods.

Table 5: Accuracy of representational effectiveness

	Number of clusters						
	20	50	80	110	140	170	200
Probabilistic clustering with FCM	**78.46%**	**82.31%**	**85.18%**	**88.69%**	**88.42%**	**89.24%**	**88.56%**
Probabilistic clustering with GMM	71.48%	82.17%	83.90%	86.52%	86.52%	85.30%	85.23%
Neural embedding based clustering [21, 40]	70.2%	78.51%	77.61%	79.65%	80.11%	79.63%	80.34%
Co-occurrence based clustering [32]	66.80%	66.40%	67.54%	68.72%	68.75%	69.02%	68.45%
Topic vector	60.61%	64.20%	63.64%	63.80%	63.74%	62.90%	62.90%
Doc2Vec	72.47%						
Word2Vec averaging	70.91%						
Bag-of-words	65.9%						
LSA	68.67%						

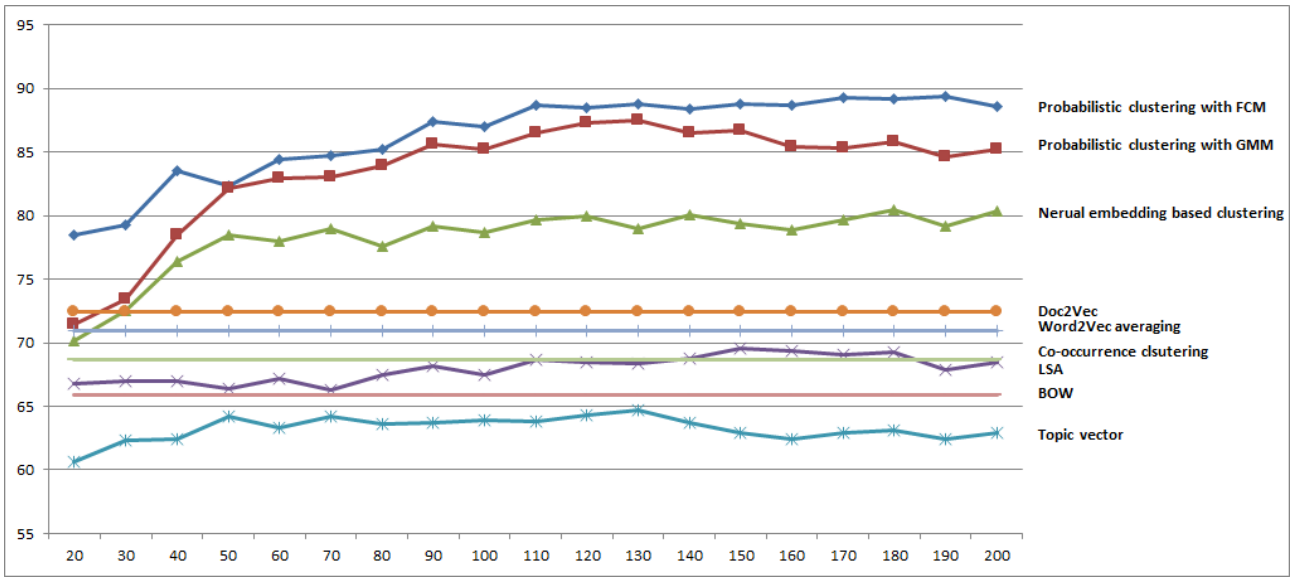


Figure 3: Accuracy of representational effectiveness

4.3.2 Classification performance

Table 6 and Figure 4 show the results of classification performance of customer-voice data with respect to varying dimensions. As observed, the experimental results are similar to those of representational effectiveness. The probabilistic word clustering based approach with FCM method outperforms all other document representation methods in all dimensions in a manner similar to the experiment of representational effectiveness.

Table 6: Accuracy of classification performance

	Number of clusters						
	20	50	80	110	140	170	200
Probabilistic clustering with FCM	**79.50%**	**83.51%**	**85.34%**	**84.92%**	**86.51%**	**87.76%**	**86.3%**
Probabilistic clustering with GMM	77.22%	80.46%	80.51%	83.92%	80.3%	82.17%	83.45%
Neural embedding based clustering [21, 40]	75.24%	79.19%	78.78%	79.40%	79.12%	80.83%	80.94%
Co-occurrence based clustering [32]	64.15%	65.97%	66.42%	67.42%	65.45%	64.81%	63.71%
Topic vector	59.00%	56.26%	63.20%	66.44%	64.60%	66.81%	66.76%
Doc2Vec	72.22%						
Word2Vec averaging	68.56%						
Bag-of-words	64.67%						
LSA	65.43%						

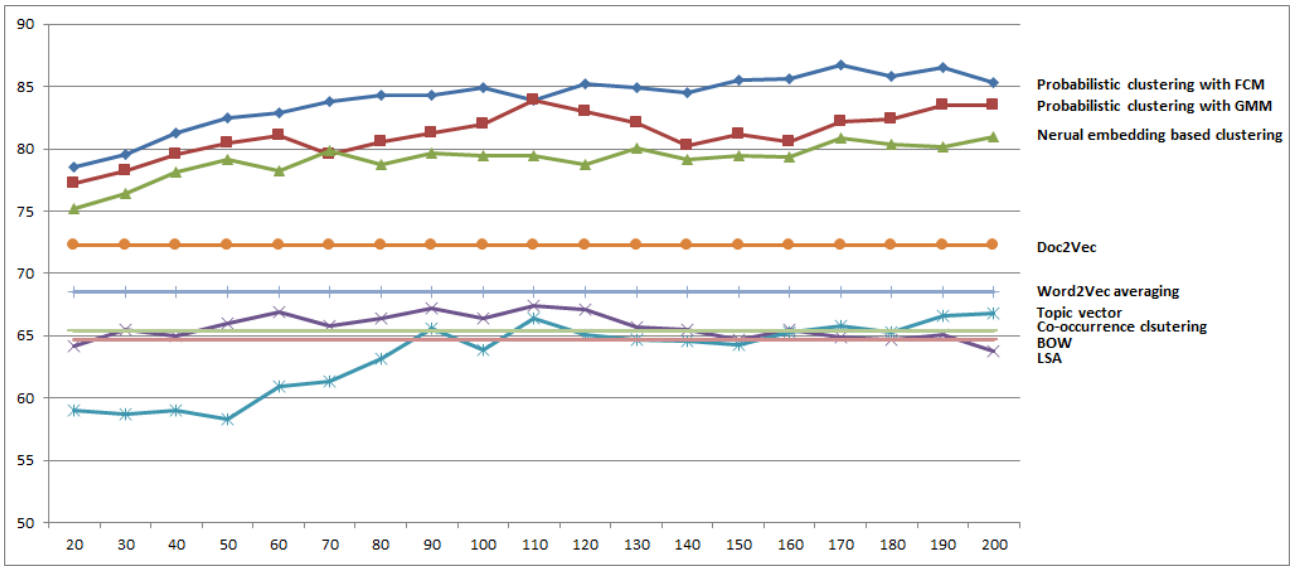


Figure 4: Accuracy of classification performance

4.4 Representational interpretation

This study provides an intuitive interpretation for the generated vector. Thus, the fore-mentioned approach is considered as a suitable method for the classification of customer-voice data. The strength of the approach is inherited by the approach proposed in the present study. Table 7 shows that the proposed method successfully offers a clear interpretation of the generated vector. The words in the cluster as shown in table 7 indicates that each cluster contains words that are closely related to each class. This implies that the customer-voice data in each class are represented by words in frequent clusters and the name or topic can be easily assigned to each cluster.

Table 7: Example of representation interpretation

customer-voice example	Class	Most frequent cluster	Words in most frequent cluster (Distance to centroid)
there are other problems, the most of which involves display brightness...	Display	3rd / 70	*Screen*(0.7934), *Brightness*(0.8024), *Display*(0.9036)
When I took pictures, they were saved in the sd card until now...	Camera & Gallery	24th / 70	*Camera*(0.8023), *Photo*(0.8903), *Shutter*(0.9724)
After mounting camplus, i press shutter button, then the backup battery is activated...	Battery & Charging	12th / 70	*Battery*(0.7236), *Charge*(0.7438), *Charging*(0.8523)
I got various accessories as well as G5. Then the VR device leave much to be desired...	Accessory	52th / 70	*Accessory*(0.7224), *Toneplus*(0.9042), *VR*(1.0035)

4.5 Application in Reuter dataset

In order to show the objective results, we carried out same experiment on public available data, Reuter dataset. Reuter dataset consists of 24,000 randomly selected articles form Reuter websits. These articles are labeled by Reuter website into 8 different classes as Entertainment, Sports, Technology, Market, Politics, Business, World and Health.

Figure 5 and Figure 6 show the results of representational effectiveness and classification performance of Reuter dataset. It shows the similar results with customer-voice dataset. Our proposed methods outperform all other document representation methods in all dimensions

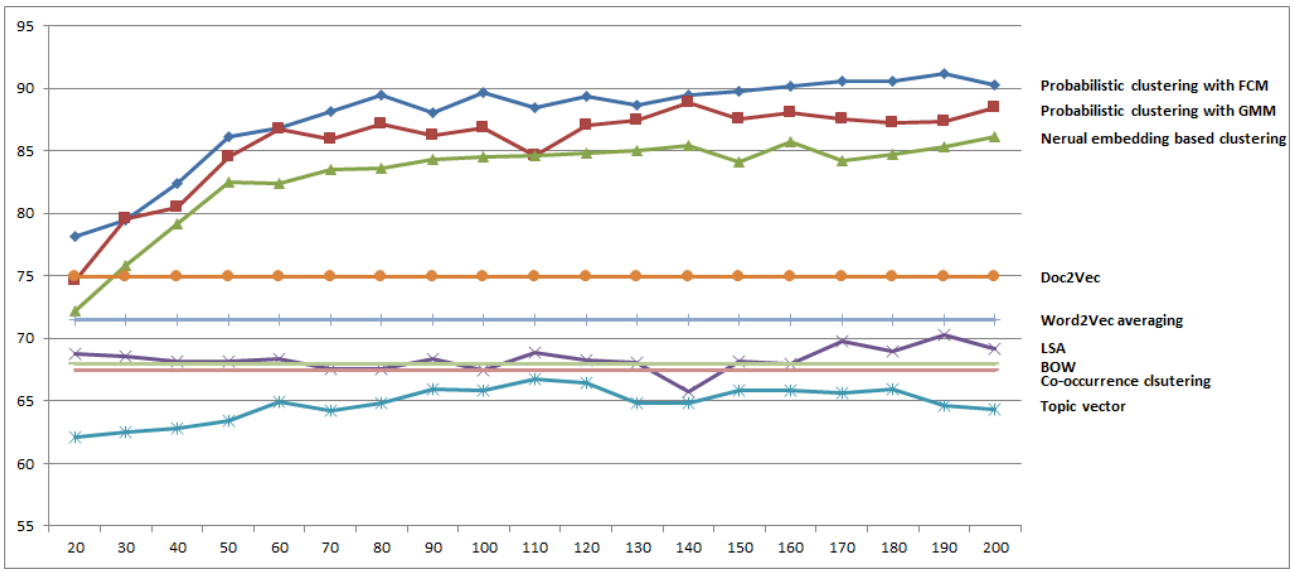


Figure 5: Accuracy of representational effectiveness of Reuter dataset

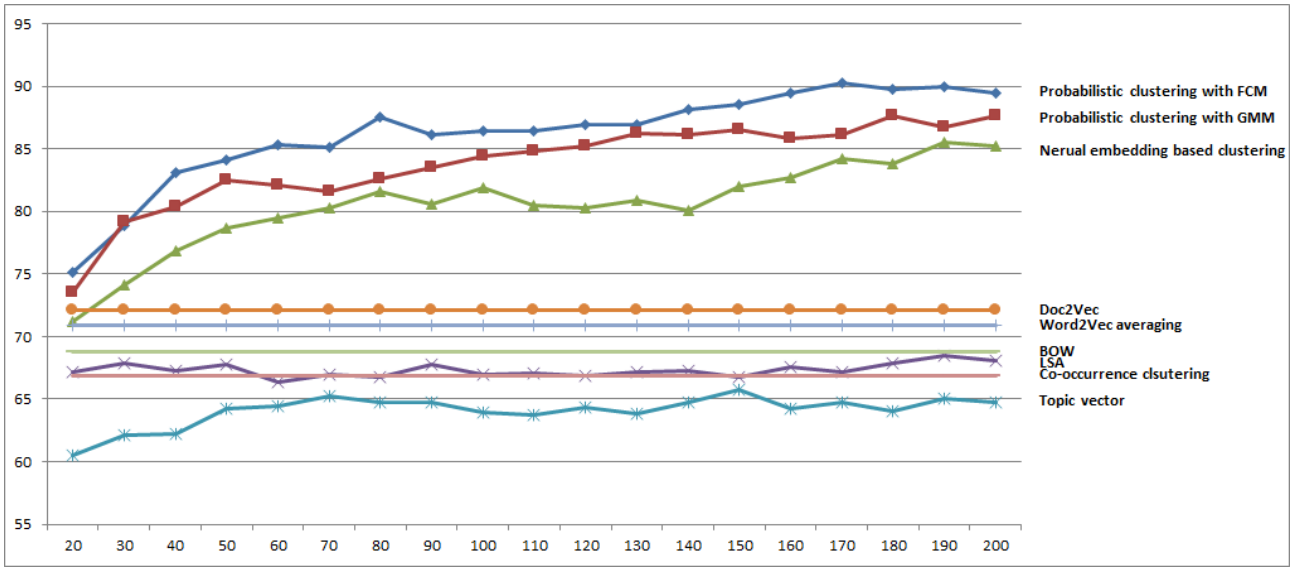


Figure 6: Accuracy of classification performance of Reuter dataset

5 Conclusion

This study involved the construction of an automatic classifier for customer-voice data using a newly proposed document representation method. The proposed method is referred to as the probabilistic word clustering based approach. There are many studies based on word clustering in text mining field, however, the previous studies as discussed in extant studies could not reflect the membership strength of each word with each cluster since it was based on a hard clustering method. Hence, in the present study, a probabilistic word clustering based approach based on the probabilistic clustering method considering membership strength of each word with each cluster was proposed. It is expected that the proposed method is robust with respect to customer-voice data that is composed of extremely unstructured texts even including typos by considering the membership strength of those words.

The proposed method outperforms all other document representation methods in actual experiments on representational effectiveness and classification performance. The proposed method achieved an accuracy of 89.24% with respect to representational effectiveness and an accuracy of 87.76% with respect to classification performance. The number of classes corresponds to 12, and this corresponds to a considerably higher absolute performance level. The reason for the higher performance of the proposed method is that it captures membership strength of words with respect to each cluster. Thus, it is concluded that the proposed method is an appropriate method to represent a document that is composed of unstructured texts, such as customer-voice data, that is to be applied in various text mining tasks in real business by replacing previous document representation methods.

In this study, the membership strength of each word is considered to construct a robust representation of an unstructured text. Future works could explore preliminary ways to remove problematic words, such as typos or other informal terms, to make a more robust representation of an unstructured text. It is expected that a more accurate classifier can be constructed by removing the preliminary problematic words and simultaneously considering the membership strength of words. Additionally, in the present study, performance of the proposed method is compared to unordered document representation methods including bag-of-words, doc2vec, LSA and bag-of-concepts approaches to focus on demonstrating the advantage of considering the membership strength. Future research will also involve a comparison of the proposed method to ordered document representation methods, such as convolutional neural networks based model [11, 22], to extend the scope of the study. Finally, the impacts of various soft clustering methods in addition to the fuzzy C-means and Gaussian mixture model that are used in this study will also be compared to build a more accurate classifier. Furthermore, it is expected that the fuzzy bag-of-concepts approach proposed in the present study after an in-depth analysis can be widely applied in various text mining tasks in real business environments.

6 reference

References

[1] Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al. *Modern information retrieval*, volume 463. ACM press New York, 1999.

[2] Ron Bekkerman, Ran El-Yaniv, Naftali Tishby, and Yoad Winter. Distributional word clusters vs. words for text categorization. *Journal of Machine Learning Research*, 3(Mar):1183–1208, 2003.

[3] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[4] Hafida Bouziane, Belhadri Messabih, and Abdallah Chouarfia. Profiles and majority voting-based ensemble method for protein secondary structure prediction. *Evolutionary bioinformatics online*, 7:171, 2011.

[5] Lijuan Cai and Thomas Hofmann. Text categorization by boosting automatically extracted concepts. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 182–189. ACM, 2003.

[6] Zhiqiang Cai, Xiangen Hu, Haiying Li, and Art Graesser. Can word probabilities from lda be simply added up to represent documents? In *Proceedings of the 9th International Conference on Educational Data Mining*, 2016.

[7] Scott Cost and Steven Salzberg. A weighted nearest neighbor algorithm for learning with symbolic features. *Machine learning*, 10(1):57–78, 1993.

[8] Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27, 1967.

[9] Andrew M Dai, Christopher Olah, and Quoc V Le. Document embedding with paragraph vectors. *arXiv preprint arXiv:1507.07998*, 2015.

[10] Pedro Domingos and Michael Pazzani. On the optimality of the simple bayesian classifier under zero-one loss. *Machine learning*, 29(2-3):103–130, 1997.

[11] Cícero Nogueira dos Santos and Maira Gatti. Deep convolutional neural networks for sentiment analysis of short texts. In *COLING*, pages 69–78, 2014.

[12] Susan T Dumais. Latent semantic analysis. *Annual review of information science and technology*, 38(1):188–230, 2004.

[13] Mario A. T. Figueiredo and Anil K. Jain. Unsupervised learning of finite mixture models. *IEEE Transactions on pattern analysis and machine intelligence*, 24(3):381–396, 2002.

[14] Stephen I Gallant. *Neural network learning and expert systems*. MIT press, 1993.

[15] Steven P Gaskin, Abbie Griffin, John R Hauser, Gerald M Katz, and Robert L Klein. Voice of the customer. *Wiley International Encyclopedia of Marketing*, 2010.

[16] Masood Ghayoomi. Word clustering for persian statistical parsing. In *Advances in Natural Language Processing*, pages 126–137. Springer, 2012.

[17] Abbie Griffin and John R Hauser. The voice of the customer. *Marketing science*, 12(1):1–27, 1993.

[18] Zellig S Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

[19] C Bedzek James. Pattern recognition with fuzzy objective function algorithms. *Kluwer Academic Publishers*, 1981.

[20] Gerald M Katz. The one right way to gather the voice of the customer. *PDMA Visions Magazine*, 25(4):1–6, 2001.

[21] Han Kyul Kim, Hyunjoong Kim, and Sungzoon Cho. Bag-of-concepts: Comprehending document representation through clustering words in distributed representation. *Neurocomputing*, 2017.

[22] Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

[23] Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. Recurrent convolutional neural networks for text classification. In *AAAI*, pages 2267–2273, 2015.

[24] Thomas K Landauer, Peter W Foltz, and Darrell Laham. An introduction to latent semantic analysis. *Discourse processes*, 25(2-3):259–284, 1998.

[25] Pat Langley, Wayne Iba, and Kevin Thompson. An analysis of bayesian classifiers. In *Aaai*, volume 90, pages 223–228, 1992.

[26] Quoc V Le and Tomas Mikolov. Distributed representations of sentences and documents. In *ICML*, volume 14, pages 1188–1196, 2014.

[27] David D Lewis. Naive (bayes) at forty: The independence assumption in information retrieval. In *European conference on machine learning*, pages 4–15. Springer, 1998.

[28] Christopher D Manning and Hinrich Schütze. *Foundations of statistical natural language processing*, volume 999. MIT Press, 1999.

[29] Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *Bulletin of mathematical biology*, 52(1-2):99–115, 1990.

[30] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[31] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[32] Olga Mitrofanova. Automatic word clustering in studying semantic structure of texts. *Advances in Computational Linguistics: Research in Computing Science. Mexico*, 41:27–34, 2009.

[33] Antonio Mucherino, Petraq J Papajorgji, and Panos M Pardalos. k-nearest neighbor classification. In *Data Mining in Agriculture*, pages 83–106. Springer, 2009.

[34] Carlos Orrite, Mario Rodríguez, Francisco Martínez, and Michael Fairhurst. Classifier ensemble generation for the majority vote rule. In *Iberoamerican Congress on Pattern Recognition*, pages 340–347. Springer, 2008.

[35] Kenji Sagae and Andrew S Gordon. Clustering words by syntactic similarity improves dependency parsing of predicate-argument structures. In *Proceedings of the 11th International Conference on Parsing Technologies*, pages 192–201. Association for Computational Linguistics, 2009.

[36] Sujan Kumar Saha, Pabitra Mitra, and Sudeshna Sarkar. Word clustering and word selection based feature reduction for maxent based hindi ner. In *ACL*, pages 488–495, 2008.

[37] Magnus Sahlgren. *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. PhD thesis, Institutionen för lingvistik, 2006.

[38] S Fouzia Sayeedunnissa, Adnan Rashid Hussain, and Mohd Abdul Hameed. Supervised opinion mining of social network data using a bag-of-words approach on the cloud. In *Proceedings of Seventh International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA 2012)*, pages 299–309. Springer, 2013.

[39] Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

[40] Vıctor Suárez-Paniagua, Isabel Segura-Bedmar, and P Martínez. Word embedding clustering for disease named entity recognition. In *Proceedings of the Fifth BioCreative Challenge Evaluation Workshop*, pages 299–304, 2015.

[41] Bruce D Temkin, Bob Chatham, and Michelle Amato. The customer experience value chain: An enterprisewide approach for meeting customer needs. *Forrester Research. March*, 15, 2005.

[42] Vladimir Naumovich Vapnik and Vlamimir Vapnik. *Statistical learning theory*, volume 1. Wiley New York, 1998.

[43] Vapnik N Vladimir and V Vapnik. The nature of statistical learning theory, 1995.

[44] Strother H Walker and David B Duncan. Estimation of the probability of an event as a function of several independent variables. *Biometrika*, 54(1-2):167–179, 1967.

[45] Chao Xing, Dong Wang, Xuewei Zhang, and Chao Liu. Document classification with distributions of word vectors. In *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*, pages 1–5. IEEE, 2014.

[46] Chao Xing, Dong Wang, Xuewei Zhang, and Chao Liu. Document classification with distributions of word vectors. In *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*, pages 1–5. IEEE, 2014.

[47] Shi Zhong. Efficient online spherical k-means clustering. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 5, pages 3180–3185. IEEE, 2005.