Intelligent RFQ Summarization Using Natural Language Processing, Text Mining, and Machine Learning Techniques

Amy J. C. Trappey, National Tsing Hua University, Taiwan
https://orcid.org/0000-0001-7651-7012

Ai-Che Chang, Shih Hsin University, Taiwan*
https://orcid.org/0000-0002-1524-5225

Charles V. Trappey, National Yang Ming Chiao Tung University, Taiwan
https://orcid.org/0000-0002-3069-6702

Jack Y. C. Chang Chien, National Tsing Hua University, Taiwan

ABSTRACT

Request for quotation (RFQ) is a lengthy document soliciting vendor products and services according to rigid specifications. This research develops an integrated natural language processing (NLP), text mining, and machine learning approach for intelligent RFQ summarization. Over 1,300 power transformer RFQ requests are used to build a word-embedding model for training and testing. Domain keywords are extracted using N-gram TF-IDF. The method automatically extracts essential specifications such as voltage, capacity, and impedance from RFQs using text analytics. The K-means algorithm groups the sentences of each specification. The TextRank algorithm identifies important sentences of all specifications to generate RFQ summaries. The summarization system helps engineers shorten the time to identify all specifications and reduces the risk of missing important requirements during manual RFQ reading. The system helps improve the complex product design for manufacturers and improve the cost estimation and competitiveness of quotations in a highly competitive marketplace.

KEYWORDS

Automatic Summarization, K-Means Clustering, Key Term Extraction, Natural Language Process (NLP), Request for Quotation (RFQ), Text Mining

1. INTRODUCTION

The rapid growth of research, development, and publication of technical documents in any given domain is causing information overload in public and private sectors. In the business world, a request for quotation (RFQ) is an invitation for competitive bids, issued by a corporation to invite suppliers or contractors to submit their bids for products or services, where the product and service requirements are specified. Functional specifications, payment terms, quality level per item, and contract length are critical aspects of RFQs. An RFQ is often confusing in terms of the technical writing within a
given knowledge domain and has strict requirements for design specifications, process technology, and standard compliances. Manufacturers wanting to participate in the bidding process must consider all specifications in the RFQ before submitting their bids in a timely manner (usually within seven days). Bidders must quickly estimate the costs of all required components and production processes to provide the most attractive quotation which wins the bid and brings increased profits and repeat business and service.

The RFQ is often used in highly-customized industries. The more complex the product manufacturing process, the more lengthy the RFQ text. The length of a power transformer RFQ document often exceeds fifty pages. However, a quick review of inquiries, accurate extraction of important information, and timely return of quotations are important for winning high value procurement bids. Traditionally, the interpretation of RFQs relies on senior engineers and domain experts. If negligent misquotation occurs, the costs are underestimated and the profits compromised. With the advancement of natural language processing (NLP) and text mining techniques, machines can assist engineers and save considerable effort summarizing the RFQ specifications and accurately estimate costs. The end result helps transformer manufacturers increase the chances of winning bids with reasonable profit margins and insures repeat business by satisfying fundamental needs quickly and efficiently.

The key terminologies (key terms) and sentence patterns in domain-specific RFQs which are similar are taken into consideration when developing an intelligent summarization system. Automatic text-based summarization is an accepted methodology for generating machine-generated summaries (Neto et al., 2002; Nenkova & McKeown, 2012; Lloret & Palomar, 2012). Automatic summarization is the process of shortening a set of data and text computationally to create a summary that represents the most important information in the original content (Saggion & Poibeau, 2013). The purpose of this research is to develop the integrated NLP, text mining, and machine learning methodology for automatic and intelligent summarization of RFQ technical documents. The paper provides a detailed structure of the proposed RFQ summarization methodology using key term extraction, summary generation, and critical requirement verification. Key term extraction using TF-IDF, n-gram, and Word2vec provide the experiment verification data for the methodology. The study collects over 1,300 RFQs for key term extraction and model training with forty additional testing RFQs used to evaluate the Compression Ratio (CR) and Retention Ratio (RR) statistics. Enterprises can train the intelligent RFQ summarization system with the machine learning approach to process a large number of existing RFQ documents which support the best bidding, design, and production strategies.

The paper is organized as follows. Section 2 reviews and discusses the related work. The three-phase modules of summary generation and benchmark methodologies are proposed and described in Section 3. Section 4 provides the experimental outcome of automatic summarization using an in-use collection of engineering documents. The case provides validation of the applied methodology. Concluding remarks for the research development and implementation are depicted in Section 5.

2. BACKGROUND AND RELATED WORK

The literature related to text mining using key term extraction, word embedding, automatic text summarization, clustering, and knowledge ontology of power transformer design and manufacturing are included in this section for review and discussion.

2.1 Key Term Extraction

Keyword extraction is an important task in the field of text mining (Siddiqi & Sharan, 2015), which is the initial and essential step in this research. Key term extraction filters input documents and selects important sentences consisting of the key terms. Siddiqi and Sharan (2015) noted that key phrase extraction is divided into four methods including rule based linguistic analysis, statistics, machine learning, and domain specific approaches. The rule based linguistic approach is derived from linguistic
knowledge and features, such as lexical analysis, syntactic analysis, and discourse analysis. These methods require computation and domain knowledge to obtain accurate results (Ercan & Cicekli, 2007).

The statistical approach is based on the linguistic corpus and quantitative analysis of the data in the corpus. One advantage of statistical analysis is that it can be applied to different languages but with less accuracy than linguistic approaches. The term frequency and inverse document frequency (TF-IDF) is one of the classical statistical methods proposed by Salton et al. (1983). TF-IDF uses the occurrence counts of terms (words and phrases) in the text corpus to evaluate the importance of terms. Domain specific approaches exploit the professional knowledge in various specific domain corpora and extract the key terms in given domain. C-value and NC-value are often used for extracting key terms in a specific domain (Frantzi & Ananiadou, 1999). C-value is a domain specific method for automatic term recognition. Wu et al. (2005) proposed a key terms extraction approach called key phrase identification program (KIP) which uses human-identified samples to assign weights to candidate key terms.

The TextRank algorithm is used to find the order of sentences with high importance and extract the sentences with the highest meaning in ranked order to generate the summarization (Mihalcea, 2004). The concept of TextRank originates from PageRank (PR) which is an algorithm used by Google Search to rank web pages in their search engine results. PageRank measures the importance of website pages (Page et al., 1999). The PageRank algorithm outputs a probability distribution used to represent the likelihood that a person randomly clicking on links will arrive at any particular page. The relationships between pages are based on the degree of connectivity between the pages and is presented as a co-occurrence matrix. The TextRank algorithm is a graph-based ranking model which determines the importance of a vertex within a graph based on the global information recursively drawn from the entire graph.

2.2 Word Embedding

Word embedding is a general description of early language models developed for large-scale processing of text based on neural nets proposed by Bengio et al. (2003). Conceptually, the models use a high-dimensional vector space where all the words are placed in a continuous vector space of a much lower dimension. Word embedding is a learned representation for text where words that have the same meaning have a similar representation. The correlation of semantic meaning of words can be expressed by their relationship in vector space. In vector space, words with similar semantic meaning are much closer to each other.

Word2vec, developed and released by Google, is a statistical method for efficiently learning standalone word embedding from a text corpus. These models are shallow, double-layer neural networks used to associate the semantic and linguistic context of words. The semantic context of a word can be defined by the adjacent words and expressed in the form of vectors by applying the learning process to a large volume of context and then mapping words into vector space. Words having similar meanings are placed closer together in vector space. With the development of highly scalable continuous bag-of-words (CBOW) and skip-gram (SG) language models for word representation learning (Mikolov et al., 2013a; 2013b), the embedding models have been shown to obtain state-of-the-art performance on many traditional language tasks after training using large text databases. Both of these models are log-linear models and use the two-step procedure for training (Mikolov et al., 2009). The main difference between CBOW and skip-gram lies in the loss function used to update the model. The CBOW model learns by predicting the current word based on its context while the skip-gram model learns by predicting the surrounding words given a word of interest.

For the skip-gram model, the target word is used for predicting the adjacent words and CBOW uses the adjacent words to predict the target word. The CBOW model’s text size is usually around five words, which is suitable for a small corpus. The accuracy rate of the CBOW model is not high, but it has low computational complexity with high speed processing. For the skip-gram model, n-Skip-gram-bi-grams can be set. The “n” refers to the size of the window, as well as the number of words
skipped or leaped over. The “bi” means the number of phrases. Skip-gram requires more training time for calculation but is more effective and accurate in expressing word semantics in vector space than CBOW (Mikolov et al., 2013b; Zhang et al., 2018). Word2vec can be used in many fields such as sentiment analysis and recommendation systems. Xue et al. (2014) used a social media text corpus as input to produce the word vectors as output. The output was used to build a sentiment dictionary using the Word2vec model and verify whether the emotions expressed were consistent. Airbnb uses skip-gram on house and apartment listings and the user’s feedback for real-time personalization when searching for similarly ranked available recommendations (Grbovic & Cheng, 2018).

2.3 Automatic Text Summarization

Keywords can be identified either manually or automatically but the former approach is time-consuming and expensive creating a need for an automated process that extracts keywords from documents. Automatic summarization is classified into extractive and abstractive techniques (Nenkova & McKeown, 2011). Extractive summarization uses text mining techniques to find the sentences which express the core meaning of a document and extract the sentences to form a summary without changing any text. The method requires a predefined model or template to create the summary (Gaikwad & Mahender, 2016). Contrasted with extractive summarization, abstractive summarization produces results closer to what people anticipate from content analysis (Nayak & Sahoo, 2018). The procedure rewords the report into a shorter format while maintaining focus and meaning (Mozhedehi et al., 2017). Abstractive summarization uses natural language processing to analyze the document and rewrite the summary without using the original sentences (Saranyamol & Sindhu, 2014; Helen, 2018).

Yogan et al. (2016) state that extractive summarization can be further classified into several subcategories including the machine learning approach, a domain specific summarization, and multi document summarization. For the machine learning approach, Naïve Bayes algorithms are an early application of machine learning. Kupiec et al. (1995) utilized Naïve Bayes to learn and derive features from data to determine the importance of sentences and extract sentences with higher ranked meaning to form a summary. Neural networks are another machine learning approach used to extract attributes from summary sentences. For domain specific summarization, the structure of the article is considered. For example, media news releases have unique structures. Domain specific summarization frameworks are used as the structure to increase the precision of the summary (Nenkova & McKeown, 2011). Multi document summarization is developed to solve information collected from various resources (Saggion & Poibeau, 2013). Cluster-based methods and graph-based methods are two popular methods used in multi document summarization (Gupta & Lehal, 2010). The cluster-based method classifies similar sentences together. Radev et al. (2004) combines the clustering method with TF-IDF for summarization. Graph-based models are used to build connections between objects. The text is regarded as a graph, sentences create a vertex, and the weight between vertexes is based on the similarity of the sentences. Hyperlink-Induced Topic Search algorithm (HITS) (Kleinberg, 1999) and PageRank (Page et al., 1999) are the most frequently referenced graph-based methods.

Abstractive summarization is classified into structured based approaches and semantic based approaches (Saranyamol & Sindhu, 2014; Gaikwad & Mahender, 2016). The structured approach identifies key information using a tree structure and template. The tree-based text in a document is referred to as a dependency tree and language generators create a summary (Kikuchi et al., 2014). A template approach represents the whole document and uses extraction rules or linguistic based rules to identify text snippets. If extracted, the snippets are placed into the template (Oya et al., 2014). The representation of semantic meaning in a document is the source of the language generation system. For multimodal semantic models and information item based methods, the semantic graph is the critical feature used for generating results. Multi-model semantic models identify the relationship among concepts which best represent the content (Greenbacker, 2011). For information item-based methods, an abstract of the source document is used to generate a summary instead of the sentences in the document (Mallett et al. 2004). The semantic graph-based method can generate Rich Semantic
Graphs (RSG), to summarize a document (Plaza et al., 2011). Extractive summarization evaluates the sentences using the features of the text or the sentences but ignores semantic importance. Our research combines Word2vec and extractive summarization to improve accuracy.

2.4 Cluster Analysis

Clustering is the unsupervised classification of patterns (observations, data items, or feature vectors) into groups (clusters) and does not require training data or pre-assumptions (Jain et al., 1999; Hosseinpour et al., 2014; Uriarte et al., 2015). Punj and Stewart (1983) added that clustering seeks to maximize variance between groups while minimizing variance within groups. K-means clustering is the method most frequently cited for database applications and has been successfully applied to text processing (MacQueen, 1967). The objects of clustering can include data, text, and images. Runkler and Bezdek (2003) clustered the text of web pages and the sequences of web pages visited by users (weblogs). Selvam et al. (2018) used a hybrid clustering approach for improving the retrieval relevancy related to inter-personal and social event detection in a multimedia context. For helping the e-retailers differentiate themselves from the competitors, the k-means clustering technique is used to segment online shoppers (based on questionnaire responses) into four types of consumer groups (Prashar et al., 2019).

2.5 Ontology Schema of a Power Transformer

The purpose of an electric power transformer is to increase (step-up) voltage for transmission or decrease (step-down) voltage for on-site home or industrial consumption. The operational principle of a transformer is based on electromagnetic induction (Smith, 2014). Electromagnetic induction is the process of generating current using a magnetic field. A magnetic field and electric conductor (e.g., a coil of wire) move relative to one another. Electricity may increase or decrease voltage when the current runs through the electromagnetic induction system. Transformers are most commonly used for increasing Alternating Current (AC) voltages (a step-up transformer) or decreasing AC voltages (a step-down transformer) for electric power applications and for coupling the stages of signal-processing circuits. Transformers can also be used for voltage isolation, where the voltage in equals the voltage out, with separate coils not electrically bonded to one another. The structure of transformers can be separated into internal and external constructs. The internal construct includes bushings, a cooling system, capacity, winding wire, insulating oil, and tap changers (De Rybel et al., 2009). The external construct consists of assembled accessories such as oil temperature indicators, winding temperature indicators, Buchholz relays, and gas detectors.

An ontology schema is used to represent knowledge of a particular domain by constructing a graphic-based semantic diagram to depict the relationships between knowledge concepts (Kozlowski & Rybinski, 2017). When exploring the competitive environment of a technology-oriented domain, a visual ontology can be constructed using automatic key term extraction from the document corpus (Trappey et al., 2019b). The hierarchical Latent Dirichlet Allocation (LDA)-based approach is a computationally intelligent method that can be used to automatically discover topics and their frequently used terms from a set of technical domain documents such as patents, request for quotations, and international manufacturing standards. Thus, a top-down ontology graphical map or a semantic schema can be generated to represent the collective knowledge of the given domain (Trappey et al., 2021). In this study’s case domain, the ontology schema of electric power transformers consists of three main parts: material, connectivity, and function (Trappey et al., 2019a), as illustrated in Figure 1. A classical large-size power transformer is shown in Figure 2.
3. METHODOLOGY

This research proposes a computer-supported system to summarize the essential specification
requirements from power transformer RFQs. The methodology is divided into three modules including key term extraction, summary generation, and key requirement verification. Key term extraction and the Word2vec model training are used after domain specific RFQ documents are collected and pre-processed. The Word2vec model, a machine learning (ML) neural network model, is trained to convert words into vector space for the knowledge domain (i.e., power transformer technology). The key term database and the Word2vec model are applied to identify the key sentences for document summary. The TextRank algorithm is applied to rank the important sentences in a given RFQ. For key requirement verification, the clustering algorithm is adopted to analyze the customers’ frequently asked requirements or specifications. The key phrases and sentences of each cluster are interpreted and explained. The main modules of the proposed RFQ summarization system are depicted in the following subsections and the modules’ process flow is shown in Figure 3.

3.1 Data Pre-processing
The power transformer RFQs are collected as data inputs for the study. Since power transformers are highly-customized, the content of the RFQ is complicated and has rigid requirements for design specifications, process technology, and international industrial standards. Before key term extraction, pre-processing steps are used to improve accuracy. The pre-processing step includes word tokenization, lemmatization, and stop words removal. The natural language toolkit (NLTK) is the package provided by the Python 3.7 version for natural language processing tasks. Word_tokenize separates the sentence into words. Lemmatization reduces the inflected words to ensure that the root word belongs to the language (Hardeniya, 2015). Low impact words and phrases with low meaning (e.g., ‘the’, ‘is’, ‘of’, ‘which’, and ‘on’) are deleted and all words are converted to lowercase.
3.2 Key Term Extraction

Text mining techniques are used to extract key phrases from the transformer RFQs. To retrieve the sentences based on the key terms and remove unnecessary information, the TF-IDF and n-gram methods are used to facilitate the result of automatic summary generation. TF-IDF proposed by Salton and McGill (1983) is a commonly used information retrieval technique used to determine the importance of a word in a set of corpora. TF calculates the occurrence frequencies of terms across an RFQ, while IDF measures the rarity of the term in the corpus of RFQs. This research uses TF-IDF to identify the top ranked key terms and establish the TF-IDF matrix of the RFQs. The scikit-learn suite in Python is used to calculate the TF-IDF value.

The n-gram model is used to complement TF-IDF. In the transformer knowledge domain, many terminologies are composed of two or more words (e.g., oil temperature indicator). N-gram represents a contiguous sequence of n items from a given text or speech (Koehn, 2009; Nguyen, 2016). An n-gram of one item, two items, or three items is referred to as a “unigram,” a “bigram,” or a “trigram,” respectively. English cardinal numbers are sometimes used for four-gram and five-gram. Using these n-grams and the probabilities of the occurrences of certain words in sequence improves the prediction of auto-completion systems. The research sets n to be two to five because n is unknown. The NLTK suite in Python is applied to extract n-grams key phrases. The discovered key terms are verified and corrected by the domain expert.

3.3 Word2vec Model Training

Word2vec is a machine learning approach that generates word vectors from a training text dataset. The research converts the text into a matrix for conducting and training using the Word2vec model and Gensim suite in Python. Word2vec is a combination of CBOW and the skip-gram model which learns weights that act as word vector representations. The parameters of the Gensim Word2vec model training are shown in Table 1.

| Parameter | Explanation |
|-----------|-------------|
| size      | Dimensionality of the word vectors and the default is 100. |
| window    | The maximum distance between a target word and words around the target word. The default window is 5. |
| min_count | Words with occurrence less than this count will be ignored. The default for the minimum count is 5. |
| workers   | The number of partitions during training and the default is 3. |
| sg        | The training algorithm, either CBOW (0) or skip-gram (1). The default training algorithm is CBOW. |

Parameter size is the dimension of the word vector and the dimension greatly influences the importance of the word vector. If the dimension is too small or too large, the relationship between the words is difficult to represent. This research sets 300 dimensions according to the research verified as the most complete expression of word meaning (Pennington et al., 2014). The parameter window is set to five which means the algorithm uses five characters before and after the target word for prediction. The parameter min_count is set to ten to allow words with occurrence of less than ten to be omitted from the training algorithm. The parameter workers uses the default setting and the parameter sg uses the value one. As mentioned in Section 2, the skip-gram model can learn better representations for infrequent words than CBOW. For the features of the transformer RFQ, the study
adopts skip-gram to discover the lower frequency phrases but the highly important specification of the RFQ (e.g., burr, a kind of cut steel sheet). For calculating the distance between each sentence, the text is tokenized into a combination of terms. The study averages the dimension of each word vector to equalize the vector length of each sentence.

### 3.4 Summary Generation

The additional RFQ corpus is used as new data to train the Word2vec model and extract the key terms. Word_tokenize in the Python NLTK suite is used to tokenize the sentence into a combination of terms and extract the important sentences based on the key term database generated by the training RFQs. The TextRank algorithm is used to find the order of sentences with high importance and extract the sentences with the higher ranking to generate the summarization (Mihalcea, 2004).

The concept of TextRank originates from PageRank (PR), an algorithm used by Google to rank web pages from search engine results. PageRank measures the importance of website pages (Page et al., 1999). PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites (Google.com). A PageRank results from a mathematical algorithm based on a webgraph created by all World Wide Web pages as nodes and hyperlinks as edges. Every page has some number of forward links (outedges) and backlinks (inedges). The PageRank algorithm outputs a probability distribution used to represent the likelihood that a person randomly clicking on links will arrive at any particular page. The relationships between pages are based on the degrees of connectivity between the pages, which can also be presented as a co-occurrence matrix. Figure 4 and Figure 5 shows an example of PageRank probabilities and their corresponding co-occurrence matrix.

![Figure 4. An example PageRank probability distribution](image)
After constructing the co-occurrence matrix weights, the PR value of each page is calculated. PageRank integrates the incoming and outgoing links into a single model and the formula to compute the PR value of a web page is expressed as:

$$PR(V_i) = (1 - d) + d \times \sum_{j : \text{In}(V_j)} \frac{1}{|\text{Out}(V_j)|} PR(V_j)$$  \hspace{1cm} (1)

Where $d$ is a damping factor that is set between 0 and 1 and is typically set to 0.85. For a given vertex $V_i$, let $\text{In}(V_i)$ be the set of vertices that point to predecessors, and let $\text{Out}(V_j)$ be the set of vertices $V_i$ that point to successors. The TextRank algorithm is a graph-based ranking model which weighs the importance of a vertex within a graph based on the global information recursively drawn from the entire graph. It is useful to indicate and incorporate into the model to measure the strength of the connection between two vertices $V_i$ and $V_j$ as a weight $w_{ji}$ added to the corresponding edge that connects the two vertices. The formula is follows.

$$WS(v_i) = (1 - d) + d \times \sum_{v_j \in \text{In}(V_i)} \frac{w_{ji}}{\sum_{v_k \in \text{Out}(V_j)} w_{jk}} WS(V_j)$$  \hspace{1cm} (2)

In comparison with Eq. (1) of PageRank, Eq. (2) adds a weighted item $w_{ji}$, which represents the different degree of importance between the two vertices, $V_i$ and $V_j$. Sentence extraction ranks the sentences and a vertex is added to the graph for each sentence in the text. To establish connections.
(edges) between sentences, the study defines a similarity relation where similarity is measured as a function of content overlap. If there are \( n \) sentences, \( k \times n \) graphs are constructed and the similarity of two sentences are calculated using the similarity formula:

\[
\text{Similarity}(S_i, S_j) = \frac{|\{w_k \mid w_k \in S_i, w_k \in S_j\}|}{\log|S_i| + \log|S_j|} (3)
\]

\( S_i \) and \( S_j \) represent the sentences \( i \) and \( j \), and \( w_k \) represents the terms that appear in both sentences. We divide by the logarithm of the number of terms in the two sentences and add them. As the connectivity of the graph increases (i.e. larger number of edges), convergence is achieved after a few iterations and the similarity between the two sentences can be calculated. The resulting graph is highly connected, with a weight associated with each edge, indicating the strength of the connections between various sentence pairs in the text (Mihalcea, 2004).

The sentence vector generated by Word2vec overcomes the sparseness problem in traditional text representations and contains semantic information between terms. Therefore, the study uses this semantic information to weight the transfer probability between sentence graph nodes from TextRank. A document can be clustered into several sub-clusters according to the similarity between sentence vectors. The farther a term is from the centroid of the subcluster, the more it reflects the different information within the subcluster. The study clusters these terms by calculating the cosine similarity \( \theta \) between any two sentences using the formula below.

\[
\text{Similarity}(A, B) = \cos \theta = \frac{A \times B}{AB} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2 \times \sum_{i=1}^{n} B_i^2}} (4)
\]

where \( A, B \) respectively represent the two sentences and \( \text{Similarity}(A, B) \) indicates the similarity between them including semantic relevance. The value of the cosine similarity is between -1 to 1. If the value is 1, the angle between the two vectors is 0, then these sentences are identical. The value -1 represents two vectors pointing in opposite directions, which means these two sentences are different. A high cosine similarity value represents sentences with highly similar semantics and increases their importance.

The process of summary generation follows four steps:

1. Tokenize each sentence and construct the graph based on each sentence (vertex).
2. Calculate the value of the cosine similarity.
3. Calculate the importance of each sentence via iterations.
4. Extract the \( n \) important sentences to generate the summarization

### 3.5 Evaluation of the Summary and Model

To evaluate the validity of the generated summary, the compression ratio (CR) and retention ratio (RR) are used as indicators to measure the output and optimize the Word2vec model (Hovy & Lin, 1999). CR refers to the ratio of the number of sentences in the summary to the number of sentences in the original document. RR measures the volume of information retained in the generated summary. An automatic fill-in summary form is developed to quantify the information volume. The study inputs the original document and the generated summary into the form respectively. Each sentence with the
key term will be fill a cell of the form based on the specification of key phrases. Counting the number of the cells and the RR value can be calculated. The formulas are shown as follows.

\[
\text{Compression Ratio} = \frac{\text{Number of sentences in generated summary}}{\text{Number of sentences in original document}} \tag{5}
\]

\[
\text{Retention Ratio} = \frac{\text{Number of cells in summary table of generated summary}}{\text{Number of cells in summary table of original document}} \tag{6}
\]

This research used the three training datasets to train the Word2vec model, and forty additional RFQs are used to test the model performance for CR and RR to find the optimized model.

### 3.6 Sentence Clustering and Key Requirement Verification

Voltage, capacity, and impedance are the most important specifications in the knowledge domain for electric power transformers. One type of transformer has various specifications, and one specification has many requirements. To discover common requirements, this research clusters the sentences of the three essential specifications using the K-means algorithm and extracts the key terms from each cluster respectively. The 1,331 training RFQs are used for clustering and the forty additional testing RFQs are used for verification. The sentences based on the three essential specifications are extracted, vectorizing, and optimized using the Word2vec model. The distance between the clustering centers of each group is calculated and placed next to the nearest cluster so that the study can interpret the features of each cluster based on the extracted key terms. The elbow method is one of the most popular methods used to select the optimal number of clusters by fitting the model with a range of values for \( k \) in the K-means algorithm. The elbow method requires drawing a line plot between the sum of squared errors (SSE) value on the Y axis and the number of clusters on the X axis and finding the point representing the inflection point. The formula is shown as follows:

\[
\text{SSE} = \sum_{i=1}^{k} \sum_{p \in C_i} (p - m_i)^2 \tag{7}
\]

\( C_i \) is the \( i \)th cluster, \( p \) is an object in cluster \( i \), \( m_i \) is the center of cluster \( i \). The research sets \( k \) from 1 to 19 to find out the optimal number of clusters and extracts the key terms under each cluster.

### 4. EXPERIMENT: CASE ANALYSIS AND MODEL VERIFICATION

The collected 1,331 precedent transformer RFQs are used to extract the key terms and train the domain Word2vec model for summary generation. The summarization case study is conducted by training and testing the model and algorithms.

#### 4.1 Key Term Extraction

After data pre-processing, text mining techniques (TF-IDF and n-gram) are used to extract key phrases from the 1,331 transformer RFQs that contain over 20 million words. This research uses TF-IDF to identify the top ranked key terms and establish the TF-IDF matrix of the RFQs. The scikit-learn suite in Python is used to calculate the TF-IDF value. The n-gram model is used to complement TF-IDF since more than two words (not single words) are often used to describe the knowledge domain.
of transformers. The research sets n to be 2 to 5 and the NLTK suite in Python is applied to extract n-grams key phrases. The top 100 key terms of each method are extracted, validated, and supplemented by the domain expert and the partial key terms of each method are shown in Table 2.

Table 2. The partial key terms of each method

| Method   | Key terms                                                                                                                                 |
|----------|-------------------------------------------------------------------------------------------------------------------------------------------|
| TF-IDF   | voltage, hv, tap, purchaser, tank, lv, spare part, generator, onaf, rat, mva, pressure, attachment, valve, onan, circuit drawing, terminal, oltc, air, ground, oil |
| Two-gram | code standard, technical specification, surge arrester, generator transformer, test inspection, auxiliary transformer, technical requirement, temperature rise, draw document, circuit breaker, ambient temperature, operation maintenance |
| Three-gram | technical specification transformer, site test commission, transformer onan onaf, applicable code standard, neutral ground resistor, installation site test, steam turbine generator, voltage primary secondary, onan onaf onaf, combine cycle plant |
| Four-gram | oman tank terminal company, valley frwy hermitage date, frwy hermitage date page, item quot deadl deliv, carolina dominion fortune shenango, page item quot deadl, hermitage date page item, date page item quot, version specification date please, version document duplicate berkshire |
| Five-gram | item quot deadl deliv date, dominion north carolina dominion fortune, dominion virginium dominion north carolina, hermitage date page item quot, date page item quot deadl, specification recent version document duplicate, specification date please consult online, recent version document duplicate berkshire |

In Table 2, the key terms of TF-IDF are the most common words used to represent knowledge related to transformers (e.g., oil means the insulation oil in the transformer, tank is the outer shell of the transformer). The key terms of the multi-grams present a more complete meaning than the extracted words of TF-IDF because most terminologies in the transformer domain consist of two or three words. Some key phrases merely can be extracted by n-gram, for example, surge arrester is a device to protect electrical equipment from over-voltage caused by external (lightning) or internal (switching) events. The study sets n to three based on the results of n-gram so the meaningful key terms are retrieved by bigram and trigram. After being validated and supplemented by the domain expert, the key terms are categorized into the specification key term and the complement key term lists. The specification key term refers to the terms that clearly express the specifications and requirements of the transformer (e.g., impedance, flux density), while the complement key terms refer to words that are often used to supplement the description of the specification key terms. For instance, test and temperature are often used to indicate that the temperature of the transformer and must be tested and reported to the customer before delivery. The customer also requires the upper limit of the temperature during operation of the transformer to ensure safety and compliance to standards. The specification key term can further be classified into four categories such as electric environment, core, accessory, other requirements and complement. The total number of key terms are 415 words and the partial key terms and each number of the five categories are shown in Table 3. Both the specification keywords and the complement key terms are stored in the keyword database to analyze the sentences of the new RFQs.
4.2 Summary Generation and Evaluation

The forty additional testing RFQs are used to evaluate the validity of the Word2vec model. The TextRank algorithm and TextRank combined with Word2vec are used to evaluate the results. The required specifications of the additional RFQs vary such as the capacity may range from 780 MVA to 230MVA, voltage from 400kV to 22.5kV, and with different winding connection methods. After calculating the value of CR and RR, the study extracts the top 35% important sentences calculated by TextRank, and the result is shown in Table 4. The Number of sentences in original document means the number of sentences in the input RFQ. The Number of sentences after filtering means the number of sentences containing key terms. The Number of sentences in summary is the number of sentences in summary generated by the algorithm proposed by this research. The Amount of information in original document refers to the number of cells used for inputting the original document into the summary form. The Amount of information in summary refers to the number of cells used for the final summary document. CR is the ratio of the number of sentences in the summary document to the number of sentences in the original document, while RR is the ratio of the amount of information in the summary document compared to the amount of information in the original document. The results show the average CR is 29.33% and the average RR is 72.27%. The ratio represents that the model helps the user obtain 72.27% of the amount of information by spending 29.33% of their usual reading time. CR’s maximum percent is 33.90%, while the minimum percent is 17.10%. RR’s maximum percent is 92.98%, while the minimum percent is 15.00%. The research finds that the inconsistent wording used for the specification of power transformers is the reason for the large difference.

Table 3. The partial key terms of the five categories

| Category                | Number of key terms | Key terms                                                                 |
|-------------------------|---------------------|---------------------------------------------------------------------------|
| Electric environment    | 73                  | service class, polarity, frequency, phase, ambient temperature, capacity,  |
|                         |                     | insulation ratings, elevation, impedance, seismic                        |
| Core                    | 32                  | core, winding temperature rise, flux density, hot spot, conductor paper    |
|                         |                     | insulation, stack, current density, creep distance, winding temperature   |
|                         |                     | rise, core thickness                                                     |
| Accessory               | 144                 | oil temperature indicator, gas detector relay, oil level indicator, ir     |
|                         |                     | window, winding temperature indicator, resistance temperature detector,    |
|                         |                     | pressure relief device, surge arrester, rapid pressure rise relay, bushing |
|                         |                     | monitor                                                                  |
| Other requirements      | 19                  | weight limit, radiator, paint, cooler, paint thickness, control cabinet,  |
|                         |                     | FM, neutral resistance grounding, short circuit, dew point                |
| Complement              | 147                 | Test, loss, Temperature, ground, oil, color, IEEE, Wye, IEC, Delta        |
The combination of Word2vec and TextRank method is used to train the Word2vec model and generate the summary. The three training sets (A, B, and C) are used for model trainings and their summarization results are evaluated using the forty additional testing RFQs set.

Dataset A. The 1,331 training RFQs,
Dataset B. The 1,331 training RFQs and the 1.2 million Wikipedia text, and
Dataset C. The 1,331 training RFQs and the 1,000 transformer related papers.

The forty additional testing RFQs are used to evaluate model performance. The outputs of the three training sets that extract the top 35% important sentence are shown in Table 5. In training set A, when the average CR is 29.33%, the average RR increases from 72.27% to 80.86%, which represents the proposed method is more effective. The ratio represents the user obtains 80.86% of the amount of information by spending 29.33% of their usual reading time. In training set B, the average CR maintains the same value while the average RR declines from 80.86% to 75.82%, which shows the model had a worse effect. The content of the Wikipedia text is relatively broad, which not only focuses on the field of power transformers. Some terms with similar meanings diverge in the vector space. The study continues training dataset C, i.e., using the 1,331 training RFQs and adds 1,000 transformer related papers. The average CR is 29.33% and the average RR is 74.35%, which shows RR has not improved significantly. Most of the 1,000 transformer related papers are academic papers with words quite different from the RFQ. The RFQ content focuses on the customer’s requirements for the power transformer, while the academic paper describes the principle and discusses the operating efficiency improvement and application in the knowledge domain of electric transformers.

| RFQ 1 | RFQ 2 | RFQ 3 | ... | RFQ 40 | Average |
|-------|-------|-------|-----|--------|---------|
| Number of sentences in original document | 1,145 | 352 | 1,033 | ... | 282 |
| Number of sentences after filtering | 1,100 | 333 | 726 | 256 |
| Number of sentences in summary document | 385 | 117 | 254 | 90 |
| Amount of information in original document | 41 | 38 | 39 | ... | 31 |
| Amount of information in summary document | 33 | 22 | 32 | ... | 18 |
| CR | 33.62% | 33.24% | 24.69% | 31.91% | 29.33% |
| RR | 80.48% | 57.89% | 82.05% | ... | 58.06% | 72.27% |

The study continues training dataset C, i.e., using the 1,331 training RFQs and adds 1,000 transformer related papers. The average CR is 29.33% and the average RR is 74.35%, which shows RR has not improved significantly. Most of the 1,000 transformer related papers are academic papers with words quite different from the RFQ. The RFQ content focuses on the customer’s requirements for the power transformer, while the academic paper describes the principle and discusses the operating efficiency improvement and application in the knowledge domain of electric transformers.

Table 5 continued on next page
In summary, dataset A has the greatest effectiveness based on the highest RR value among the above experiment results. It is hypothesized that when the training set is the same type of file, the best results are obtained. After obtaining the best model, the study tests the effect of increasing the proportion of the number of extracted sentences on CR and RR. The top 45% of important sentences from the 1,331 training RFQs are extracted for testing and the results are shown in Table 6. The CR and RR comparison between Table 5 and Table 6 shows the average CR increases from 29.33% to 37.77% (an increase of 8.44%), while the average RR increase from 80.86 to 81.39% (a very small 0.53% improvement). The result represents that the user has to spend more reading time but the amount of information obtained has not been improved. The top 35% of important sentences are adopted as the optimized selection improve the system accuracy.

Table 6. The result of the extracted top 45% important sentences

|                   | RFQ 1  | RFQ 2  | RFQ 3  | ...  | RFQ 40 | Average |
|-------------------|--------|--------|--------|------|--------|---------|
| Number of sentences in original document | 1145   | 352    | 1033   | ...  | 282    |         |
| Number of sentences in summary document  | 495    | 150    | 327    | ...  | 116    |         |
| Amount of information in original document | 41     | 38     | 39     | ...  | 31     |         |
| Amount of information in summary document | 36     | 34     | 30     | ...  | 26     |         |
| CR                 | 43.23% | 42.61% | 31.66% | ...  | 41.13% | 37.77%  |
| RR                 | 87.80% | 89.47% | 76.92% | ...  | 83.87% | 81.39%  |
The study calculates the CR and RR value by extracting the top 35%, 30%, and 25% of the important sentences from the forty additional testing RFQs. Each RFQ selects the suitable percentage based on calculating the degree of decrease in the CR and RR value (Table 7). If the decrease in RR is greater than CR, a higher extraction percentage is selected; conversely, if the decrease in RR is less than CR, a smaller extraction percentage is selected. Among the 40 RFQs, 34 RFQs are selected for 35%, 2 RFQs (RFQ #6, #13) are selected for 30%, and the 4 RFQs (RFQ#21, #33, #34, #36) are selected for 25%. The result improves the top 35% of the important sentences and the optimized selection improves the accuracy of the system.

### 4.3 Key Requirement Discovery and Verification

The research clusters the sentences based on the three essential specifications (voltage, capacity, and impedance) from the 1,331 training RFQs. The sentences describing similar requirements are classified in the same group, and the key terms extracted using the K-means algorithms. The forty additional testing RFQs are used to verify the meaning of each cluster based on sentences containing critical specifications. The key terms and the related sentences are used to infer common customer requirements within each cluster. The elbow method requires drawing a line plot between the sum of squared errors (SSE) value on the Y axis and the number of clusters on the X axis and finding the point representing the inflection point. As Figure 6 given below, for n_clusters = 4 that represents the elbow start seeing diminishing returns by increasing k. The line starts looking linear. The clustering results of the three essential specifications are described in the following three subsections.

**Table 7. The result based on extracting the suitable percentage of the important sentences**

| RFQ  | Number of sentences in original document | Number of sentences in summary document | Amount of information in original document | Amount of information in summary document | Extracted percentage | CR       | RR       |
|------|------------------------------------------|-----------------------------------------|-------------------------------------------|-------------------------------------------|---------------------|----------|----------|
| 1    | 1145                                     | 385                                     | 41                                         | 35                                         | 35%                 | 33.62%   | 85.37%   |
| 2    | 352                                      | 117                                     | 38                                         | 30                                         | 35%                 | 33.11%   | 78.95%   |
| 3    | 1033                                     | 254                                     | 39                                         | 33                                         | 35%                 | 24.60%   | 84.62%   |
| 4    | 1666                                     | 413                                     | 34                                         | 31                                         | 30%                 | 24.79%   | 91.18%   |
| 5    |                                         |                                         |                                            |                                            |                     | 15.00%   | 86.67%   |
| 6    |                                         |                                         |                                            |                                            |                     | 13.00%   | 80.65%   |
| 7    |                                         |                                         |                                            |                                            |                     |          |          |
| 8    |                                         |                                         |                                            |                                            |                     |          |          |
| 9    |                                         |                                         |                                            |                                            |                     |          |          |
| 10   |                                         |                                         |                                            |                                            |                     |          |          |
| 11   |                                         |                                         |                                            |                                            |                     |          |          |
| 12   |                                         |                                         |                                            |                                            |                     |          |          |
| 13   |                                         |                                         |                                            |                                            |                     |          |          |
| 14   |                                         |                                         |                                            |                                            |                     |          |          |
| 15   |                                         |                                         |                                            |                                            |                     |          |          |
| 16   |                                         |                                         |                                            |                                            |                     |          |          |
| 17   |                                         |                                         |                                            |                                            |                     |          |          |
| 18   |                                         |                                         |                                            |                                            |                     |          |          |
| 19   |                                         |                                         |                                            |                                            |                     |          |          |
| 20   |                                         |                                         |                                            |                                            |                     |          |          |
| 21   |                                         |                                         |                                            |                                            |                     |          |          |
| 22   |                                         |                                         |                                            |                                            |                     |          |          |
| 23   |                                         |                                         |                                            |                                            |                     |          |          |
| 24   |                                         |                                         |                                            |                                            |                     |          |          |
| 25   |                                         |                                         |                                            |                                            |                     |          |          |
| 26   |                                         |                                         |                                            |                                            |                     |          |          |
| 27   |                                         |                                         |                                            |                                            |                     |          |          |
| 28   |                                         |                                         |                                            |                                            |                     |          |          |
| 29   |                                         |                                         |                                            |                                            |                     |          |          |
| 30   |                                         |                                         |                                            |                                            |                     |          |          |
| 31   |                                         |                                         |                                            |                                            |                     |          |          |
| 32   |                                         |                                         |                                            |                                            |                     |          |          |
| 33   |                                         |                                         |                                            |                                            |                     |          |          |
| 34   |                                         |                                         |                                            |                                            |                     |          |          |
| 35   |                                         |                                         |                                            |                                            |                     |          |          |
| 36   |                                         |                                         |                                            |                                            |                     |          |          |
| 37   |                                         |                                         |                                            |                                            |                     |          |          |
| 38   |                                         |                                         |                                            |                                            |                     |          |          |
| 39   |                                         |                                         |                                            |                                            |                     |          |          |
| 40   |                                         |                                         |                                            |                                            |                     |          |          |

**Figure 6. The results of the elbow method for finding the optimal number of clusters**
4.3.1 The Clustering of Voltage Related Sentences

The 4,661 voltage related sentences are clustered into 4 groups which are determined by the elbow method from the 1,331 training RFQs. The clustering result of the voltage related sentences and their key term extraction are shown in Table 8.

Table 8. The clustering result based on the voltage related sentences from the 1,331 training RFQs

| Method  | Cluster 1 (1,316 sentences) | Cluster 2 (1,246 sentences) | Cluster 3 (617 sentences) | Cluster 4 (1,482 sentences) |
|---------|----------------------------|-----------------------------|---------------------------|-----------------------------|
| TF-IDF  | kv, substation, transformer, equipment, switchgear, line, | kv, voltage, kva, transformer, rated, winding, tap, test, bushing, rating | kv, voltage, rated, bil, transformer, winding, neutral, peak | kv, transformer, power, mva, line, system, cable, bushing, protection |
| Bigram  | circuit breaker, distribution line, terminal equipment, | power transformer onan onaf, high voltage, full capacity | hertz 230, 60 hertz, 230 grdy, phase 230, voltage transformer | power transformer, onan onaf, high voltage |
| Trigram | terminal equipment olte, line terminal equipment, optical line terminal | onan onaf onaf, phase 60 hertz, onaf onaf 65, shall full capacity | hertz 230 grdy, 60 hertz 230, 230 grdy 132.8, 400 phase 230 | onan onaf onaf, phase 60 hertz, onaf onaf 65 |

The 1,249 related sentences are extracted based on the voltage key terms from the forty additional RFQs and classifies the sentences according to the key terms of the four clusters. A total of 104 sentences are categorized in Cluster 1, the term circuit and break (circuit breaker) with higher frequency represents the higher importance. The term equipment and assembly related to the transformer (e.g., switchgear) also appears frequently in the sentences. Cluster 1 represents the customer’s requirements in the assemblies of the transformer. The 1,018 sentences are classified into Cluster 2 which focuses on requirements for voltage testing, e.g., the term test, full capacity, and 60 hertz. For instance, the customer requires that the operating test conditions must be carried out at full capacity or under 60 Hz conditions. The 27 sentences grouped within Cluster 3 focus on the requirements for voltage. After reading the key sentences, almost all of them describe the rated voltage and wiring method of the transformer. The 100 sentences grouped within Cluster 4 emphasize the requirements of the circuit. The high frequency terms are bushing and cable. The bushing is an important insulation device and the cable must pass through the bushing before contacting the transformer to ensure insulation. There are many descriptions of bushings and lines based on the key sentences, e.g., customers ask that high-voltage side bushings use capacitors and bushings. The brief statement is shown in Table 9.
4.3.2 The Clustering of the Capacity Related Sentences

Capacity is another influencing parameter when designing a transformer. Capacity refers to the load that a transformer can withstand. A capacity value is obtained by multiplying voltage by current, and the unit is mainly divided into kVA and MVA. K refers to thousands, and M refers to millions, the difference between the two is 1,000 times. There are many kinds of values that appear in the RFQ based on the different phases with the different capacity algorithms. The 1,644 capacity related sentences are divided into 3 groups from the 1,331 training RFQs. The clustering result of the capacity related sentences and their key term extraction are shown in Table 10.

The 266 related sentences are extracted based on the capacity key terms from the forty additional testing RFQs. The 131 sentences grouped within Cluster 1 focus on the requirements of the capacity value (e.g., MVA, transformer, and rated). After verifying the key sentences, almost all of them describe the basic capacity requirements of the transformer including the capacity of the high voltage side circuit and the low voltage side circuit used during operation. Cluster 2 represents self-cooling...
since it has the highest frequency. Self-cooling is a cooling method, which means cooling through natural methods, without radiator or cooling oil. The description about self-cooling appears in the key sentences, e.g., the transformer should achieve the specified capacity through self-cooling. The 93 sentences grouped within Cluster 3 focus on the requirements of the cooling class. The terms Oil Natural Air Natural (ONAN) and Oil Natural Air Force (ONAF) are used to describe the cooling class of the transformer. The definition of ONAN cooling is that hot oil dissipates heat in the air by natural convection and conduction process and the oil is cooled by the circulation of natural air and passes through the radiator of the transformer. This type of cooling is used for transformer ratings up to 30 MVA. The method of ONAF cooling refers to the use of forced air to cool the transformer and is used for cooling transformer with ratings up to 60 Mega volts ampere. The brief statement is shown in Table 11.

Table 11. The customer requirement for capacity

| Cluster | Customer’s requirement | Key term | Example sentence |
|---------|------------------------|----------|------------------|
| #1 (131 sentences) | Capacity value | MVA, rated, transformer, power | • 75/100/125MVA 220/13.8kV YNd1 with NLTC ONAN/ ONAF/ONAF High  
• 10/14 MVA at 50500 Volts connected Delta. |
| #2 (42 sentences) | Self-cooling | self cool, cool rating | • Capacity self-cooled rating: 24 MVA  
• For a three-phase core-form transformer with a self-cooled rating of 5000 kVA or above… |
| #3 (93 sentences) | Cooling class | onan onaf onaf, force cooling oaf | • 30/33.6 MVA ONAN  
• Transformer rated 10MVA and above shall be rated for ONAF. |

4.3.3 The Clustering of the Impedance Related Sentences

Electrical impedance is the measure of the opposition that a circuit presents to a current when voltage is applied. %Z is a key parameter in power transformers and its value determines the available short circuit current for the secondary transformer. Impedance is commonly called %Z or % impedance. A total of 930 impedance related sentences are divided into 3 groups from the 1,331 training RFQs. The clustering result of the impedance related sentences and their key term extraction are shown in Table 12.

Table 12. The clustering result based on the impedance related sentences from the 1,331 training RFQs

| Method | Cluster 1 (473 sentences) | Cluster 2 (362 sentences) | Cluster 3 (95 sentences) |
|--------|---------------------------|---------------------------|--------------------------|
| TF-IDF | impedance, transformer, mva, voltage, high, tap, protection, system, base | impedance, voltage, tap, measurement, load, loss, transformer, shortcircuit, rated, winding | impedance, sequence, zero, measurement, voltage, test, zerosequence, positive, measured, phase |
| Bigram | phase bushing, transformer shall, locate closest, closest line, bushing locate | short circuit, impedance voltage, load loss, circuit impedance, sequence impedance | zero sequence, sequence impedance, positive zero, impedance shall, phase sequence, sequence impedance |
| Trigram | phase bushing locate, bushing locate closest, closest line follow, c400 phase bushing, audible sound level | short circuit impedance, circuit impedance load, voltage short circuit, impedance voltage short, zero sequence impedance | zero sequence impedance, sequence impedance, positive zero sequence, impedance measurement zero, phase sequence impedance |
The 178 related sentences are extracted based on the impedance key terms from the 40 additional testing RFQs. The 77 sentences grouped within Cluster 1 focus on the customer requirements for the impedance value. The description of the impedance value is expressed in percentages in the key sentences. A total of 81 sentences classified into Cluster 2 represents the requirements of short circuit impedance. Short circuit impedance refers to the impedance generated by the transformer during a short circuit. Short circuit impedance is one of the transformer performance parameters and customers ask the transformer manufacturer to verify the short circuit impedance value as requested before shipping. The 20 sentences grouped within Cluster 3 focus on the requirement of zero sequence impedance based on the key terms zero sequence and positive zero sequence. The purpose of zero sequence impedance is to analyze when the system voltage and current are asymmetrical. In a three phase system, a positive sequence set of currents produces a normal rotating field, a negative sequence set produces a field with the opposite rotation, and the zero sequence set produces a field that oscillates but does not rotate between phase windings. Customers ask the transformer manufacturer to calculate the three phase value for earth-fault protection and earth-fault current calculations. The brief statement is shown in Table 13.

Table 13. The customer requirement for impedance

| Cluster | Customer’s requirement | Key term | Example sentence |
|---------|------------------------|----------|------------------|
| #1 (77 sentences) | Impedance value | Impedance, transformer | • System impedance shall always be considered as 0%, unless agreed to in writing by NW.  
• For each option, 6% & 7%, the quotation shall state the impedance at the following five tap positions. |
| #2 (81 sentences) | Short circuit impedance | short circuit, short circuit impedance | • Three phase equivalent short circuit impedance measurement.  
• Measurement of load losses and impedance voltages / short-circuit impedances at rated frequency. |
| #3 (20 sentences) | Zero sequence impedance | zero sequence, positive zero sequence | • Measurement of zero-sequence impedance at rated frequency at principal taps and all extremes.  
• Percent impedance both positive and zero sequence between the various windings. |

4.3.4 The Customer Requirement Summary based on Clustering

According to the clustering results, the common customer requirements of the three essential specifications are summarized in Table 14. Assembly, voltage testing, voltage value, and circuit are the common requirements in the specification of voltage. Capacity has three customer requirements: capacity value, self-cooling, and cooling class. The specification of impedance has three customer requirements: impedance value, short circuit impedance, and zero sequence impedance. When an engineer receives a new RFQ, the engineer must first confirm whether the customer requirements are met using the summary customer requirements.
Table 14. Frequently used customer requirements for transformer RFQs

| Specification | Requirement #1 | Requirement #2 | Requirement #3 | Requirement #4 |
|---------------|----------------|----------------|----------------|----------------|
| Voltage       | Assembly       | Voltage testing | Voltage value  | Circuit        |
| Capacity      | Capacity value | Self-cooling    | Cooling class  |                |
| Impedance     | Impedance value| Short circuit impedance | Zero sequence impedance | |

5. CONCLUSION

For highly competitive business environments, the speed by which important information can be obtained better enables the enterprise to gain the advantage for new market opportunities. This research proposes a methodology for automatic technology extraction to summarize the text from the thousands of transformer RFQs used in the experiment. The engineering of highly customized production processes and the RFQ is regarded as pre-contract cost planning and control between the manufacturer and the customer. A pre-tender estimate is no longer an approximate estimate and is prepared by pricing the bills of quantities ready for issuance or already issued for tendering. By checking against tenders returned for suitability, the two parties use the RFQ to check whether the cost of the pricing is within the previous estimate and budget. The enterprise prices the bill of materials before the formal issue of tender documents or before the return of tenders.

The research uses a machine learning approach and algorithms to develop an automatic collective intelligence summarization system applicable to any given domain for RFQ document corpora. Machine learning assists engineers to reduce the probability of misjudgment and saves considerable reading time. The transformer manufacturers increase the chances of winning bids with reasonable remuneration. The engineering document automatic summarization system helps the engineer shorten the time to obtain the key information of RFQs and reduce the risk of missing important information during manual reading. The result demonstrates that the model helps the user obtain 80% of the most critical information by spending 30% of their usual reading time. The discovered customer common requirements based on cluster analysis helps the engineer to avoid missing essential requirements.

The study uses advanced data analytics and machine learning to provide strategic and decision support for competitive and increased profit bidding. The developed engineering knowledge platform serves as the basis for the next generation of intelligent manufacturing services to ensure sustainable global competitiveness. Improving product design and testing, cost estimation, and quotation accuracy are the future research topics for the field. In order to obtain accurate results, the automatic summarization system and research process must be developed in collaboration with the domain experts since intangible domain knowledge is problematic for the algorithms to capture with sufficient accuracy.

ACKNOWLEDGMENT

This research is partially supported by the Ministry of Science and Technology and Ministry of Education (Taiwan) individual research grants (Grant numbers: MOST-110-2221-E-007-113-MY3, MOST-108-2221-E-007-075-MY3, and MOE-10101-A-P11005).
REFERENCES

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research, 3*(Feb), 1137–1155.

De Rybel, T., Singh, A., Vandermaar, J. A., Wang, M., Martí, J. R., & Srivastava, K. D. (2009). Apparatus for online power transformer winding monitoring using bushing tap injection. *IEEE Transactions on Power Delivery, 24*(3), 996–1003. doi:10.1109/TPWRD.2009.2022674

Ercan, G., & Cicekli, I. (2007). Using lexical chains for keyword extraction. *Information Processing & Management, 43*(6), 1705–1714. doi:10.1016/j.ipm.2007.01.015

Frantzi, K. T., & Ananiadou, S. (1999). The C-value/NC-value domain-independent method for multi-word term extraction. *Journal of Natural Language Processing, 6*(3), 145–179. doi:10.5715/jnlp.6.3_145

Gaikwad, D. K., & Mahender, C. N. (2016). A review paper on text summarization. *International Journal of Advanced Research in Computer and Communication Engineering, 5*(3), 154–160. doi:10.17148/IJARCCE.2016.5340

Grbovic, M., & Cheng, H. (2018). Real-time personalization using embeddings for search ranking at Airbnb. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 311-320*. doi:10.1145/3219819.3219885

Greenbacker, C. F. (2009). Automated text summarization and the summarist. In I. Mani & M. Maybury (Eds.), *Advances in Automatic Text Summarization* (pp. 1–14). Academic Press.

Honey, E., & Lin, C. Y. (1999). Automated text summarization and the summarist. In I. Mani & M. Maybury (Eds.). Advances in Automatic Text Summarization (pp. 1–14). Academic Press.

Hosseinpour, F., Amoli, P. V., Farahnakian, F., Plosila, J., & Häälääinen, T. (2014). Artificial immune system-based intrusion detection: Innate immunity using an unsupervised learning approach. *International Journal of Digital Content Technology and its Applications, 8*(5), 1-12.

Hovy, E., & Lin, C. Y. (1999). Automated text summarization and the summarist. In I. Mani & M. Maybury (Eds.). Advances in Automatic Text Summarization (pp. 1–14). Academic Press.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys, 31*(3), 264–323. doi:10.1145/331499.331504

Kikuchi, Y., Hirota, T., Takamura, H., Okumura, M., & Nagata, M. (2014). Single document summarization based on nested tree structure. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 315-320). doi:10.3115/v1/P14-2052

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the Association for Computing Machinery, 46*(5), 604–632. doi:10.1145/324133.324140

Koehn, P. (2009). *Statistical machine translation*. Cambridge University Press. doi:10.1017/CBO9780511815829

Kozlowski, M., & Rybinski, H. (2017). Word Sense Induction with closed frequent termsets. *Computational Intelligence, 33*(3), 335–367. doi:10.1111/cin.12091

Kupiec, J., Pedersen, J., & Chen, F. (1995). A trainable document summarizer. *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 68-73.

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. *Proceedings of the International conference on machine learning, 1188-1196.*

Lloret, E., & Palomar, M. (2012). Text summarisation in progress: A literature review. *Artificial Intelligence Review, 37*(1), 1–41. doi:10.1007/s10462-011-9216-z
MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability (pp. 281-297). Berkeley, CA: University of California Press.

Mallett, D., Elding, J., & Nascimento, M. A. (2004). Information-content based sentence extraction for text summarization. Proceedings of International Conference on Information Technology: Coding and Computing, ITCC 2004, 2, 214-218. doi:10.1109/ITCC.2004.1286634

Mihalcea, R. (2004). Graph-based ranking algorithms for sentence extraction, applied to text summarization. Proceedings of the ACL Interactive Poster and Demonstration Sessions, 170-173. doi:10.3115/1219044.1219064

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013a). Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems, 3111–3119.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013b). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Mikolov, T., Kopecky, J., Burget, L., Glembek, O., & Cernocky, J. (2009). Neural network based language models for highly inflective languages. In Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, (pp. 4725-4728). IEEE. doi:10.1109/ICASSP.2009.4960686

Mozhedehi, A. T., & Rahimi, S. R. (2017). An Overview on Extractive Summarization. IEEE International Conference on Knowledge based Engineering and Innovation (KBEI).

Nenkova, A., & McKeown, K. (2012). A survey of text summarization techniques. In C. Aggarwal & C. Zhai (Eds.), Mining text data (pp. 43–76). Springer. doi:10.1007/978-1-4614-3223-4_3

Nenkova, A., & McKeown, K. (2011). Automatic summarization. Foundations and Trends® in Information Retrieval, 5(2–3), 103-233. 10.1561/1500000015

Neto, J. L., Freitas, A. A., & Kaestner, C. A. A. (2002). Automatic Text Summarization Using a Machine Learning Approach. In G. Bittencourt & G. L. Ramalho (Eds.), Advances in artificial intelligence (pp. 205–215). Springer. doi:10.1007/978-3-540-36127-8_20

Nguyen, V. H., Nguyen, H. T., Duong, H. N., & Snasel, V. (2016). N-gram-based text compression. Computational Intelligence and Neuroscience, 2016, 9483646. Advance online publication. doi:10.1155/2016/9483646 PMID:27965708

Oya, T., Mehdad, Y., Carennini, G., & Ng, R. (2014). A template-based abstractive meeting summarization: Leveraging summary and source text relationships. Proceedings of the 8th International Natural Language Generation Conference (INLG), 45-53. doi:10.3115/v1/W14-4407

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The PageRank citation ranking: bringing order to the web. Technical Report. Stanford InfoLab. Retrieved from http://ilpubs.stanford.edu:8090/4022/

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 1532-1543. doi:10.3115/v1/D14-1162

Plaza, L., Diaz, A., & Gervás, P. (2011). A semantic graph-based approach to biomedical summarization. Artificial Intelligence in Medicine, 53(1), 1–14. doi:10.1016/j.artmed.2011.06.005 PMID:21752612

Prashar, S., Tata, S. V., Parsad, C., Banerjee, A., Sahakari, N., & Chatterjee, S. (2019). Clustering e-shoppers on the basis of shopping values and web characteristics. Journal of Global Information Management, 27(2), 24–38. doi:10.4018/JGIM.2019040102

Punj, G., & Stewart, D. W. (1983). Cluster analysis in marketing research: Review and suggestions for application. JMR, Journal of Marketing Research, 20(2), 134–148. doi:10.1177/002224378302000204

Radev, D. R., Jing, H., Styš, M., & Tam, D. (2004). Centroid-based summarization of multiple documents. Information Processing & Management, 40(6), 919–938.

Runkler, T. A., & Bezdek, J. C. (2003). Web mining with relational clustering. International Journal of Approximate Reasoning, 32(2-3), 217–236.
Saggion, H., & Poibeau, T. (2013). Automatic text summarization: Past, present and future. In Multi-Source, Multilingual Information Extraction and Summarization (pp. 3–21). Springer. doi:10.1007/978-3-642-28569-1_1

Sahoo, A., & Nayak, A. K. (2018). Review paper on Extractive Text Summarization. International Journal of Engineering Research in Computer Science and Engineering, 5(4), 46–52. 01.1617/vol5/iss4/pid04981

Salton, G., & McGill, M. J. (1983). Introduction to modern information retrieval. Mcgraw-Hill.

Saranyamol, C. S., & Sindhu, L. (2014). A survey on automatic text summarization. International Journal of Computer Science and Information Technologies, 5(6), 7889–7893.

Selvam, S., Balakrishnan, R., & Ramakrishnan, B. S. (2018). Ontology with hybrid clustering approach for improving the retrieval relevancy in social event detection. International Journal on Semantic Web and Information Systems, 14(4), 33–56. doi:10.4018/IJSWIS.2018100102

Siddiqi, S., & Sharan, A. (2015). Keyword and keyphrase extraction techniques: A literature review. International Journal of Computers and Applications, 109(2), 18–23. doi:10.5120/19161-0607

Smith, R. T., Taylor, S., & Maher, S. (2014). Modelling electromagnetic induction via accelerated electron motion. Journal of Physics, 93(7), 802–806.

Trappey, A. J., Trappey, C. V., & Govindarajan, U. H. (2019a). Knowledge Extraction of RfQ Engineering Documents for Smart Manufacturing. In 22th International Conference Advances in Materials and Processing Techniques, Taipei, Taiwan.

Trappey, A. J. C., Trappey, C. V., Chang, A.-C., & Li, J. X. K. (2019b). Deriving competitive foresight using an ontology-based patent roadmap and valuation analysis. International Journal on Semantic Web and Information Systems, 15(2), 68–91. doi:10.4018/IJSWIS.2019040104

Trappey, C. V., Chang, A.-C., & Trappey, A. J. C. (2021). Building an internet-based knowledge ontology for trademark protection. Journal of Global Information Management, 29(1), 123–144. doi:10.4018/JGIM.2021010107

Uriarte, R. B., Tsaftris, S., & Tiezzi, F. (2015). Service clustering for autonomic clouds using random forest. In Proceedings of the 2015 IEEE/ACM 15th International Symposium on Cluster, Cloud, and Grid Computing (CCGrid) (pp. 515-524). doi:10.1109/CCGrid.2015.41

Wu, Y. F. B., Li, Q., Bot, R., & Chen, X. (2005). Domain-specific keyphrase extraction. In Proceedings of the 14th ACM international conference on Information and knowledge management (pp. 283-284). ACM.

Xue, B., Fu, C., & Shaobin, Z. (2014). A study on sentiment computing and classification of Sina Weibo with Word2vec. In IEEE International Congress on Big Data (pp. 358-363). IEEE. doi:10.1109/BigData. Congress.2014.59

Yogan, J. K., Goh, O. S., Halizah, B., Ngo, H. C., & Puspalata, C. (2016). A review on automatic text summarization approaches. Journal of Computational Science, 12(4), 178–190. doi:10.3844/jcssp.2016.178.190

Zhang, C., Wang, X., Yu, S., & Wang, Y. (2018). Research on keyword extraction of Word2vec model in Chinese corpus. In 2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS) (pp. 339-343). IEEE.
Amy J. C. Trappey received her Ph.D. in Industrial Engineering from Purdue University. Prof. Trappey’s research interests are in knowledge engineering and intelligent systems, particularly for IP and patent analytics, e-business, and e-manufacturing. Prof. Trappey is currently serving as the Associate Editor for Advanced Engineering Informatics and IEEE Transactions on SMC: Systems. Dr. Trappey is an ASME and ISEAM Fellow and a senior member of IEEE.

Ada Chang received her Ph.D. in Industrial and Business Management from National Taipei University of Technology, Taiwan. Her research interests are in technology innovation, data driven marketing, and IP issues related to marketing. Dr. Chang has ten years of teaching experience and working experience focusing on promoting technology innovation.

Charles V. Trappey received his Ph.D. in Consumer Science and Retailing from Purdue (USA), an MS in Quantitative Business Analysis from Louisiana State University (USA), and an LLM in IP Law from Queensland University of Technology (Australia). Prof. Trappey’s research interests are in innovation and technology forecasting, IP informatics, retail business development, and technology marketing. Dr. Trappey is an Australia Trademark Attorney.

Jack Y. C. Chang Chien received his MS degree in Industrial Engineering and Engineering Management, National Tsing Hua University, Taiwan. Jack’s research focuses on automatic summarization, natural language process (NLP), and ontology schema of power transformer.