applied sciences

MDPI

Article

Predicting Project's Uncertainty Risk in the Bidding Process by Integrating Unstructured Text Data and Structured Numerical Data Using Text Mining

JeeHee Lee and June-Seong Yi *

Department of Architectural and Urban Systems Engineering, Ewha Womans University, Seoul 03760, Korea; jeehee@ewhain.net

* Correspondence: jsyi@ewha.ac.kr; Tel.: +82-2-3277-4454

Received: 29 September 2017; Accepted: 3 November 2017; Published: 6 November 2017

Abstract: The construction bidding process takes place in the early stages of a construction project. The bidding process involves many uncertainties since the construction risk factors are predicted and reflected in the bid price before construction commences. Therefore, bidders should thoroughly understand the uncertainty factors of the project before make bidding decisions. The uncertainty risk of construction projects is determined by the content that is contained in the bidding document. If the information provided in the bidding document is not accurate and is unclear, the uncertainty of the projects increases. Thus, this study is performed to predict risks in the bidding process of construction projects by analyzing the uncertainty of the bidding document and using it as factors to predict a project's bidding risk. To achieve this, bidding risk prediction modeling was conducted using the pre-bid clarification information of each project. In addition, text mining on pre-bid RFI documents, which are in an unstructured text data format, was performed and the results of text mining were used as major influencing factors for the risk prediction models. As a result, the accuracy of the risk prediction model including text data was improved (72.92%) when compared to the prediction model using only numeric data (52.08%). The results of this study are expected to strengthen the possibility of further similar studies in the future since it enhances the predictive accuracy by incorporating the uncertainty of the bidding document, which is rarely considered in previous studies.

Keywords: construction bidding risk; pre-bid clarification; text mining; risk prediction modeling; unstructured text data

1. Introduction

Construction companies take into account the characteristics, profitability, competitiveness, and risk factors of a construction project when determining a bid price for construction projects [1–5]. However, construction projects have various uncertainties in the bidding phase since the bidding phase is the earliest stage of the project, and the difficulties and risk factors of the construction need to be predicted and reflected in the bid price before construction commences. Therefore, deciding the appropriate bid price is difficult because bidders need to thoroughly understand the uncertainty factors of the project before making bidding decisions. If the bid price is too high, it is difficult to award the project, but if the bid price is too low, even if the bidder awards the project, the profitability can decrease. Therefore, deciding the appropriate bid price is one of the most important decision making tasks in the bidding process.

In general, it is almost impossible to identify all of the potential risk factors of a project within a short bidding period and reflect them in the bid price. Since information uncertainty is high at the beginning of the project, it is difficult to perform specific risk assessment considering the project characteristics.

The bidding document provided by the project owner is the most representative information available to the bidders during the bidding stage. The bidding document includes owner's project planning intention and the overall construction requirements. In particular, the uncertainty of the project is determined by the information contained in the bidding document, which is also directly related to the contractual issue. Thus, a very thorough review and risk analysis in the bidding document can lead to a successful project. The information provided in the bidding document, however, is not always clear and adequate. The drawings, specifications, and contract conditions that are included in the bidding document, contain a number of uncertainties, such as ambiguity, missing information, incorrect information, and discrepancies between information [6,7]. These uncertainties can then lead to overestimation or underestimation in construction bidding [8]. Moreover, in some recent international projects change orders may not be considered when the bidders fail to properly review the errors, omissions, and discrepancies in the bidding document during the bidding process [9].

In many previous studies, the risk factors that should be considered in the bidding stage have been analyzed and risk prediction models have been developed to estimate the appropriate bid price. However, most of these studies have focused on a project's external influencing factors, such as the competition level of bidding and the bidder's experience of similar projects, or have quantified the risks based on the results of questionnaires of experts. In most studies, the project-oriented characteristics, technical level of the project, and inherent uncertain risks, which affect and determine a project's risk level have not been analyzed enough.

In this study, we analyze the text data of pre-bid clarification information in the bidding stage in order to identify the type of uncertainty risks that occurs mainly in the bidding document. The text analysis results were selected as the main factor of the bid risk prediction model in order to render the bidding risk prediction more effective. This study is performed to confirm that the accuracy of risk prediction can be improved by including unstructured data analysis when compared to predicting bid risks using only numeric data. In other words, this study confirms whether the accuracy of the risk prediction model can be improved by incorporating the uncertain risk factors existing in the bidding document into the bid risk prediction. To achieve this, we compare the accuracy of the risk prediction model that integrates the unstructured data with the numeric data and the accuracy of the risk prediction model with numeric data only.

Information on the public infrastructure construction projects ordered within the last three years by the California department of transportation (Caltrans) was collected in order to analyze the bidding information including pre-bid clarification, bid price, estimated price, and the number of bidders. Based on the collected bidding information, text mining analysis was performed separately on text information, and risk prediction modeling was performed by integrating structured numeric data and unstructured text data. This study focuses on evaluating the risks that are inherent in the projects in order to derive the project-oriented bidding risk factors. The external factors, such as bidder's experiences of similar projects and the owner's risks were therefore excluded in this study.

2. Materials and Methods

2.1. Literature Review

Numerous studies have been conducted on bidding model for construction projects. Such research can be divided into the study of model development to support bid decision making [4,5,10–14] and study to determine bid cost and contingencies in consideration of bidding risks [8,15,16].

The previous studies related to the bid decision model for construction projects show the factors that affect the bid price. In [10], a quantitative analysis of the impact of the number of project bidders on project bid price was presented. The study found that reducing the number of bidders will result in increased project bid prices. In [4], a Case-based reasoning (CBR) approach in bid decision making was suggested using internal biding factors (contractor's expertise, experience, financial ability, resource, etc.) and external biding factors (nature of the work, biding requirement, social, and economic

environment). A bid decision-making model based on the fuzzy set theory was proposed in [12]. In that study, the most vital factors influencing the bid/no bid decision are the type of work, experience in similar projects, and contractual terms.

The way in which risk is accounted for in the bidding process was studied in [16]. The types of uncertainty that occur in the bidding document of construction projects were defined in [8], and the factors that cause cost overruns were identified. Based on the results, the authors of [8] proposed a bid contingency cost that can be prepared for uncertain risks.

Previous studies related to bid price forecasting and bid decision models reveal that many factors affect the bid price and bid decisions, and the project's characteristics appear to be one of the most important influential factors among many bid risk factors. While many studies have presented quantitative research results for bidding decision, the analysis of the factors that can specify the characteristics of the project (that is, the factors hidden in the bidding documents) were not usually considered. In other words, even though the bidding document contains information that describes the characteristics of the project, studies on bidding risk prediction based on detailed analysis of the bidding document have not been conducted. Most of the previous studies have analyzed data using structured numerical data, and most of these numerical data are obtained from questionnaires completed by related field experts. While the method of obtaining quantitative figures of qualitative factors through surveys is effective, this data can often be inaccurate and unreliable.

In [17], cost overrun was predicted by integrating the text data and numeric data of a construction bidding document; however, the study has a limitation since the authors did not fully consider the many potential uncertainty risks that are included in the bidding document and used only short text summary data. In addition, the prediction model has an accuracy of less than 50%, which may reduce the reliability of the analysis results.

This study considers that most of the important information that determines a project's characteristics is in the form of unstructured text data; an integrated data analysis approach using unstructured text data and numeric data is therefore proposed. In this study, we analyze the text of pre-bid clarification information in order to examine bidding document in which the characteristics of projects and risks exist, in order to identify the type of information that is not clear in the bidding document. Finally, the bidding risks of construction projects are predicted by integrating the unstructured text data with the structured numeric data. Furthermore, this study proves the effectiveness of integrated data mining by comparing the predicted results from the integrated data analysis proposed in this study and the results from only numeric data analysis.

2.2. Pre-Bid Clarification

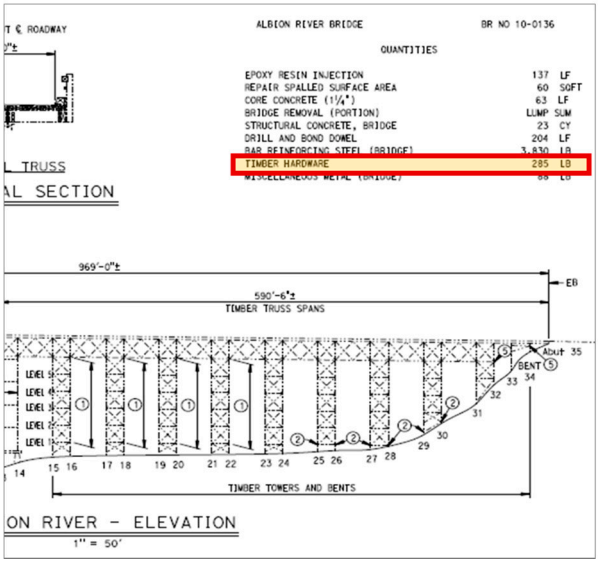
The bidding stage of all construction projects involves clarifying the owner's requirements through pre-bid meeting and pre-bid inquiry. In particular, the aim of this process is to prevent unnecessary future changes in the construction process by clearly clarifying uncertain information of the bidding document provided by the project owner. This process is called 'pre-bid clarification'. In this process, bidders submit written inquiries to the owners in accordance with the provisions of the bidding document, and the owner replies in writing to the inquirers to ensure that all of the bidders can see relevant information. This process is not merely a formal process for bidding, but is also an institutional device that can increase the accuracy of the bidder's cost estimating by clarifying uncertain and unclear information in the bidding document, thus preventing future design changes, claims, and dispute risks in advance. In addition, the pre-bid clarification information can be considered as some of the most important information that arises in the bidding stage, since it is attached to the bidding and becomes part of the contract [9]. Therefore, pre-bid clarification information is likely to include various types of potential risk factors that may arise in the bidding document. Key risk factors commonly highlighted in the bidding document could be extracted by also analyzing the pre-bid clarification information.

Figure 1 shows a sample of pre-bid clarification in infrastructure construction work ordered by Caltrans, where a bidder inquired about the discrepancy between drawings and bid item quantities,

which are part of the bidding document, resulting in the publication of an addendum. This type of pre-bid clarification information is used in this study for predicting bid risks because the risk factors and uncertainty risks of the project can be understood through pre-bid clarification information. The main information generated during the pre-bid clarification process is a pre-bid Request for Information (RFI) and an addendum that is issued as a result of the pre-bid RFI. During the pre-bid clarification period, bidders query the owner about unspecified parts of the bidding document or expected problems, and if the bidding document has errors, the owner issues an addendum. In such a case, the uncertainty of the project is partially resolved because the unclear risks have been resolved by issuing an addendum. In some cases, however, the owner does not issue addendums within the bidding period, or the owner refuses to respond to the anticipated problems, so that the bidders consider this situation as an uncertainty risk and reflects the final bid price. Thus, differences might arise between the case in which the potential risks and anticipated problems of the project were eliminated through the pre-bid clarification process and the case where potential risks were existed in the project without pre-bid clarification.

Inquiry : Bid Item 34 Timber hardware; Quantity is 1,142 LB however plan sheet 13 general plan quantities depicts 285 LB, please advise.

Response : An addendum has been issued to address this bidder inquiry. Please refer to Addendum No. 1, issued on Thursday, August 11, 2016.



Item No.	Item Code	Item Description	Unit of Measure	Estimated Quantity
21	153225	PREPARE CONCRETE BRIDGE DECK SURFACE	SQFT	53,320
22 (F)	153233	TREAT BRIDGE DECK	SQFT	53,320
23	153234	FURNISH BRIDGE DECK TREATMENT MATERIAL	GAL	593
24	153319	CORE CONCRETE (1 1/4")	LF	63
25	157560	BRIDGE REMOVAL (PORTION)	LS	LUMP SUM
26	159010	RECONSTRUCT METAL RAILING (BRIDGE)	LF	6
27	390095	REPLACE ASPHALT CONCRETE SURFACING	CY	35
28	390151	ASPHALTIC EMULSION (BONDED WEARING COURSE)	TON	2
29	395010	BONDED WEARING COURSE-GAP GRADED	TON	120
30	397005	TACK COAT	TON	0.2
31 (F)	510053	STRUCTURAL CONCRETE, BRIDGE	CY	23
32	511106	DRILL AND BOND DOWEL	LF	204
33 (F)	520102	BAR REINFORCING STEEL (BRIDGE)	LB	3,830
34 (F)	045334	TIMBER HARDWARE	LB	1,142
35 (F)	750501	MISCELLANEOUS METAL (BRIDGE)	LB	374
36	840504	4" THERMOPLASTIC TRAFFIC STRIPE	LF	12,000

Figure 1. Example of pre-bid clarification.

Figure 2 shows the distribution of overestimated projects and underestimated projects (left side histogram) and the difference between pre-bid clarified projects and unclarified projects for the collected 260 Caltrans infrastructure projects using bidder's average bid price compared to owner's estimate cost (right side box plot). The histogram in Figure 2 shows that the number of overestimated projects is somewhat higher than the number of underestimated projects. The box plot shows that the average bid price of the clarified projects is lower than that of the unclarified projects.

This result implies that the bid risk level of the projects can be predicted based on the information for pre-bid clarification. In this study, the pre-bid clarification information, which has not been discussed in many previous bidding risk studies, is used for bid risk prediction and to improve the accuracy of the prediction model. In particular, types of risks that could appear in the bidding document have been identified by analyzing pre-bid RFI text data based on the uncertainty factors of which the bidders have inquired about with the owners. In addition, information on the number of

times the addendum issued through the pre-bid RFI and the type of information that was included in the issued addendum was also used as information for predicting the project's bidding risks.

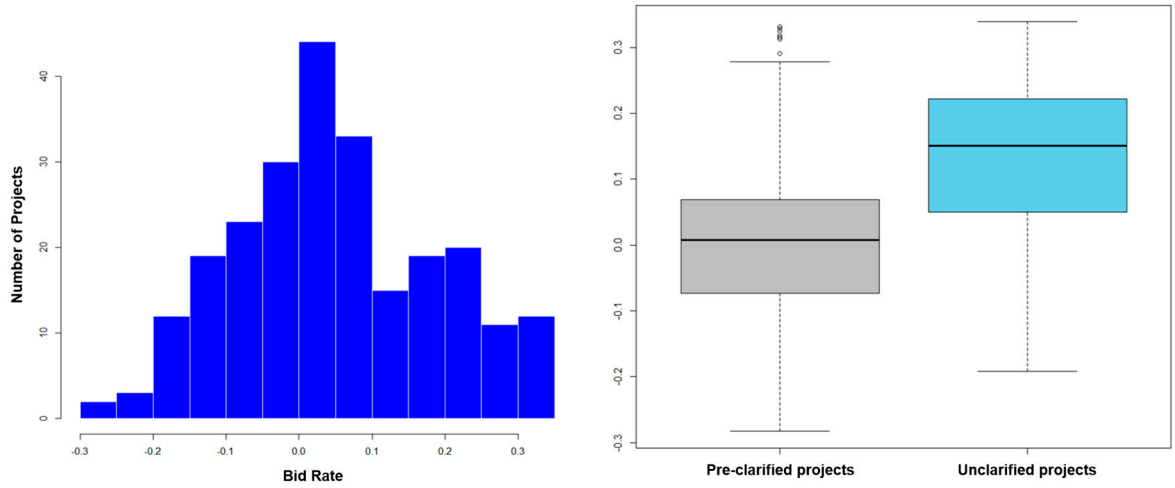


Figure 2. Distribution of bid price relative to estimated cost between pre-clarified projects and unclarified projects.

3. Materials and Method

3.1. Data Collection

260 public infrastructure projects that were ordered by Caltrans in the last three years were collected in order to predict bid risks. Caltrans orders a large number of infrastructure projects every year and it is considered that to the effective analysis of bidding information is possible because of the amount of data that is accumulated over the years. Pre-bid clarification information including pre-bid RFI, number of issued addendum, number of bidders, owner's estimate, and average bid price of 260 projects were collected. Since the pre-bid RFI is written in the text, text mining was performed to convert the unstructured text data into structured numeric data. All collected data except for the pre-bid RFI is stored as a numeric data set, and is integrated with the converted pre-bid RFI data, which is later converted into a structured format.

In this study, risk prediction modeling is performed by integrating unstructured text data and structured numeric data based on pre-bid clarification information to determine how uncertain risk factors in the bidding document affect project risk level. Therefore, only the projects of similar size and in similar locations were collected so that the factors, such as project size and location, which were not treated as risk factors in this study, would not affect the risk prediction results.

Based on the fact that the bidder's bid price as compared to the owner's estimate price in pre-clarified projects is relatively lower than that in unclarified projects, as shown in Figure 2, the bidder's average bid price when compared to the owner's estimate of each project was defined as a bid risk percentage. That is, a project with a positive bid risk percentage means that the average bid prices are higher than the owner's estimate, and a project with a negative bid risk percentage means that the average bid prices are lower than the owner's estimate. In other words, a project with a high bid risk percentage can be considered as a project with high bid risks, and a project with a low bid risk percentage can be considered a project with low bid risks. The bid risk percentage, which means the average bid price compared to the owner's estimate, is calculated using Equation (1).

$$\text{Bid Risk Percentage} = (\text{average bid price-estimate})/\text{estimate} \times 100\% \quad (1)$$

The bid risk percentage was calculated from the owner's estimates and the average bid prices of each projects, while the risk groups of all the collected projects were determined based on the bid risk

percentage value. In the previous studies that suggested the optimal bid markup when considering the risks in the bidding stage of construction projects [18,19], around 8–10% was considered reasonable to calculate the bid markup in comparison with the owner's estimate. In this study, therefore, projects with bid risk percentages exceeding 10% out of the 260 public infrastructure project data were considered to be high risk group that exceeded the appropriate level of risks. In other words, a total of 260 project data were divided into a project group with bid risk percentages that were exceeding 10% and a project group with bid risk percentages of less than 10%. To further break down risk groups, projects with bid risk percentage of 0 to 10% were allocated to a moderate risk group, while projects with less than 0% were allocated to a low risk group when considering the distribution of the collected data. The projects which an average bid cost of more than 35% when compared to the owner's estimate are excluded in this study, because it is believed such projects are incurred significant changes during the bidding process.

3.2. Modeling Process

Bid risk prediction modeling through integrated analysis of unstructured text data and structured numeric data is conducted in three stages. First, input data is refined to analyze the integrated data analysis through data collection and preprocessing. In the case of text data, unlike numeric data, a preprocessing process of segmenting sentences into tokens and converting them into vectors is additionally required. Therefore, text data and numeric data were analyzed separately during the preprocessing. The preprocessing process, which converts textual data into a structured numeric vector, is usually followed by tokenization, stopping, stemming, normalization, and vector generation [20].

After preprocessing, the text data and numeric data are integrated and ready for data analysis. Integrated data were randomly separated into a training data set and a testing data set for the learning of risk patterns. Eighty percent of the total data were used as training data and 20% as testing data in this study. The first 260 data sets collected of infrastructure projects were reduced to 243 data sets by preprocessing; 195 projects were therefore randomly allocated to the training data, and 48 projects were randomly allocated to the testing data. The three different risk group projects were allocated to the training data set and testing data set at an equal rate. Training data is learned by classification algorithms to identify risk patterns present in the data. This study applied four representative classification algorithms: ANN (Artificial Neural Network), SVM (Support Vector Machine), KNN (k-nearest neighbors), and NB (Naïve Bayes).

The classification model for each algorithm learned from training data verifies the performance of the classification models through testing data, which are not involved in the training process. Classification performances of the four different classification algorithms were compared in order to determine the classifier that has the best classification power. Moreover, classification performance with integrated data was compared to the classification performance with numeric data only to verify that the performance of the integrated prediction model is superior to the prediction model using only numeric data. Figure 3 shows the process of the prediction modeling in this study.

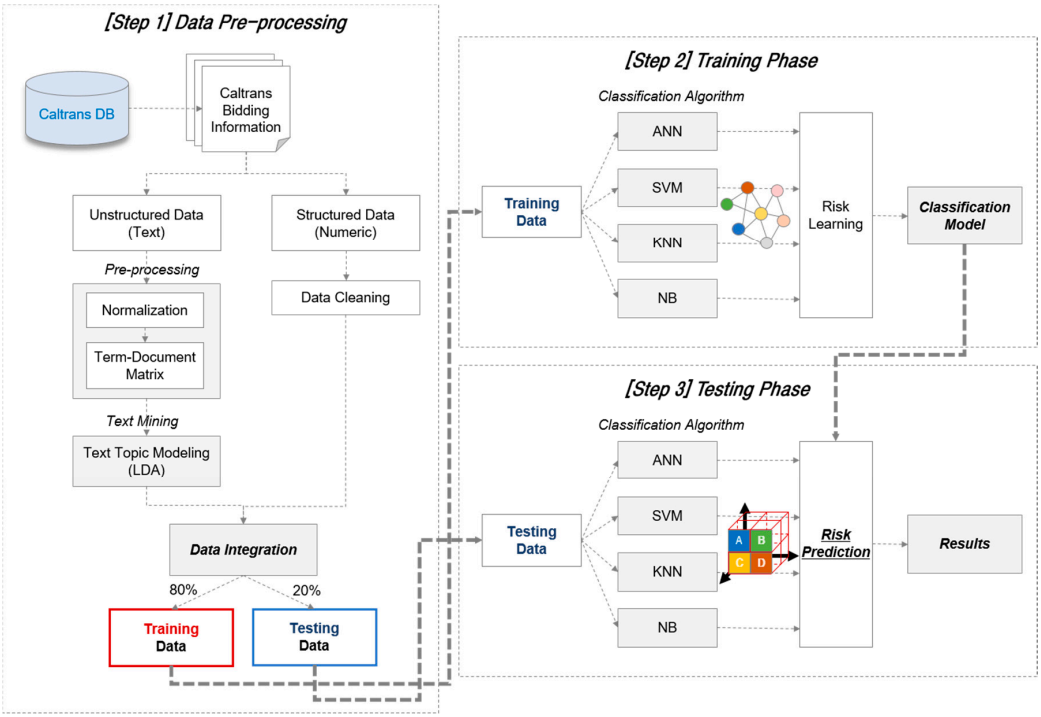


Figure 3. Research model process.

4. Data Modeling

4.1. Text Mining

4.1.1. Data Pre-Processing

After preprocessing and structuring the text data, the structured text data is integrated with the numeric data to construct an integrated data set. Of the 260 projects initially collected, text data was collected for the remaining 255 projects, excluding five projects with bid risk percentages of more than ±35%. The collected text data comprise the pre-bid RFI information of each project, including bidder's inquiries about unclear and risk information contained in the bidding document. However, to the process of manually reviewing all of the pre-bid RFI information for 255 projects is time-consuming and inefficient. Thus, text mining using R programming (ver. 3.4.1) was performed to analyze the types of pre-bid RFI information of each project.

Text mining refers generally to the process of extracting interesting and non-trivial patterns or knowledge from unstructured text documents [21]. One of the reasons for applying text mining in this study is to change the text data format to a structural vector format in order for integrated data analysis. To perform text mining on the pre-bid RFI document, appropriate preprocessing is required to ensure computers are able to automatically recognize the textual document. In other words, the unstructured information should be structured to be understood by the computers, and the unnecessary information should also be excluded. Before text preprocessing a total of 243 projects were finally selected as analysis data of the study by excluding 12 project data that do not have pre-bid RFI among the collected project data.

In the text mining, the basic structure for managing documents is the corpus, which is a set of text documents. In this study, the corpus refers to a set of 243 pre-bid RFI. The corpus can be generated from the text documents imported from a directory, vector, or data frame in R. Each document in the corpus contains unnecessary information for analysis such as punctuation, symbols, and white spaces. By removing this unnecessary information, we established a basic environment for data analysis. In addition, words that appear in the pre-bid RFI document, such as 'inquiry', 'contractor'

and 'Caltrans', do not themselves have any special meaning for analysis, but appear repeatedly in all of the documents. Thus, the words were recognized as stop words for pre-bid RFI and were eliminated for effective analysis and improvement of processing speed.

Table 1 shows the preprocessing results of text data. As a result of the preprocessing, the total number of words in the corpus decreased from 5009 terms to 3948 terms. The value indicating the sparsity of the words in the documents is 98%, indicating that only 2% of the words occur repeatedly, and 98% of the remaining words occur at a relatively low frequency.

Table 1. Preprocessing of text data.

Number of documents	243 documents
Number of terms (before preprocessing)	5009 terms
Number of terms (after preprocessing)	3948 terms
Minimum word length	3 characters
Sparsity	98%

After removing the unnecessary information of the corpus through the preprocessing process, a term-document matrix is created that converts unstructured text data into a structured form. The term-document matrix is a matrix that quantifies the frequency of words that are contained in each document and can be constructed as shown in Equation (2).

$$term - document\ matirx = \begin{pmatrix} & T_1 & T_2 & T_3 & \dots & T_t \\ D_1 & w_{11} & w_{12} & w_{13} & \dots & w_{1t} \\ D_2 & w_{21} & w_{22} & w_{23} & \dots & w_{2t} \\ D_3 & w_{31} & w_{32} & w_{33} & \dots & w_{3t} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ D_n & w_{n1} & w_{n2} & w_{n3} & \dots & w_{nt} \end{pmatrix} \quad (2)$$

However, even after preprocessing, about 4000 terms are still included in the corpus, and not all of these words have equal importance. Therefore, it is necessary to select words having a high priority according to their weight. 'Term frequency-inverse document frequency' (tf-idf) is the most representative weighting for priority of words. Tf-idf considers the frequency of words in various documents as well as the frequency of words in each document. It is based on the concept that as the frequency of the occurrence of a particular word increase in a document, the importance of that word decrease [22]. Tf-idf can be calculated using the following formula.

$$w_{t,d} = \log(1 + tf) \times \log N/df \quad (3)$$

$$w_{t,d} : TF - IDF\ weighting$$

$$tf : Term\ Frequency$$

$$N : Total\ Number\ of\ Documents$$

$$df : Document\ Frequency$$

The mean value of tf-idf for all terms in the corpus was 0.0442. Thus, very low weighted terms were excluded from the analysis by removing terms with tf-idf value less than 0.0442. The number of words in the corpus was reduced to 1566 after applying tf-idf weighting.

4.1.2. Text Topic Modeling

To understand the types of information included in the pre-bid RFI documents, text topic modeling, which is frequently used in text mining, was performed. Text topic modeling is based on the idea that documents are mixtures of topics, where a topic is a probability distribution over words [23]. Text topic modeling is a way to identify topics that cover each document by grouping the documents

according to the occurrence probability of words. Text topic modeling is distinguished from the text clustering method since one document can contain several topics. The Latent Dirichlet Allocation (LDA) method was applied as a topic modeling algorithm. LDA is an unsupervised generative probabilistic model for collections of discrete data, such as text corpora. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics [24]. Assuming that several topics are mixed in each document, the results of learning can be applied to the new document to analyze the topic of the new document. An inference process is required to analyze the topic of given documents through the LDA model. Gibbs sampling, which is easy to provide a relatively efficient method of extracting a set of topics from large corpus [25–27], was used for the inference in this study. 11 topics were derived from the topic modeling of the pre-bid RFI document using the LDA algorithm, as shown in Figure 4. In topic modeling, trial and error is repeated several times in the process of determining the number of topics. In this study, 11 topics are finally obtained using the harmonic mean method [28] to determine the optimal number of topics. The scalar value of the Dirichlet hyper-parameter for topic proportions, alpha, was set at 0.02 and the number of iterations was 2000. Figure 4 shows a bar-chart of the probability of the topic terms representing 11 topics.

Figure 4. Results of pre-bid Request for Information (RFI) topic modeling.

Figure 5 shows a visualization of the topic modeling results using LDAvis, one of the R packages. LDAvis gives a global view of the topics (and the way in which they differ from each other), while at the same time allowing for a deep inspection of the terms most strongly associated with each individual topic [29]. The left side of Figure 5 shows the distribution of each topic, and the right side shows a bar chart of the key terms that represent the selected topic (see topic 5 in Figure 5). In particular, the blue bar refers to the global frequency of each word and the red bar refers to the local frequency in the bar chart of the right view. For example, the first word 'asphalt' in the bar chart of Figure 5 shows that most of the blue bar is occupied by the red bar, implying that the word is rarely used in other topics except topic 5, and the word can thus be considered as a representative key word of topic 5.

The 11 topics that were found as a result of the topic modeling were examined using the LDAvis visualization tool, and the title of each topic could be deduced as follows:

- Topic 1: Surface roughness requirements
- Topic 2: Disadvantaged Business Enterprise (DBE) work category

- Topic 3: Guardrail system detail
- Topic 4: Excavation quantity and payment
- Topic 5: Highway planting and irrigation
- Topic 6: Hot Mixed Asphalt (HMA) spec.
- Topic 7: Type of bid item and quantity clarification
- Topic 8: Manufacturing and supplying of equipment
- Topic 9: Monitoring report and duties
- Topic 10: Working day and schedule
- Topic 11: Temporary traffic control

LDA topic modeling not only allowed us to understand the type of information handled by each document, but also provided numeric data, which is the probability of occurrence for each project topic. This numeric data obtained from text analysis are used as a parameter for predicting bid risks of infrastructure construction, and the final data set is composed by integrating the results of the text analysis with other numeric data, such as the number of bidder, the number of issued addenda, etc.

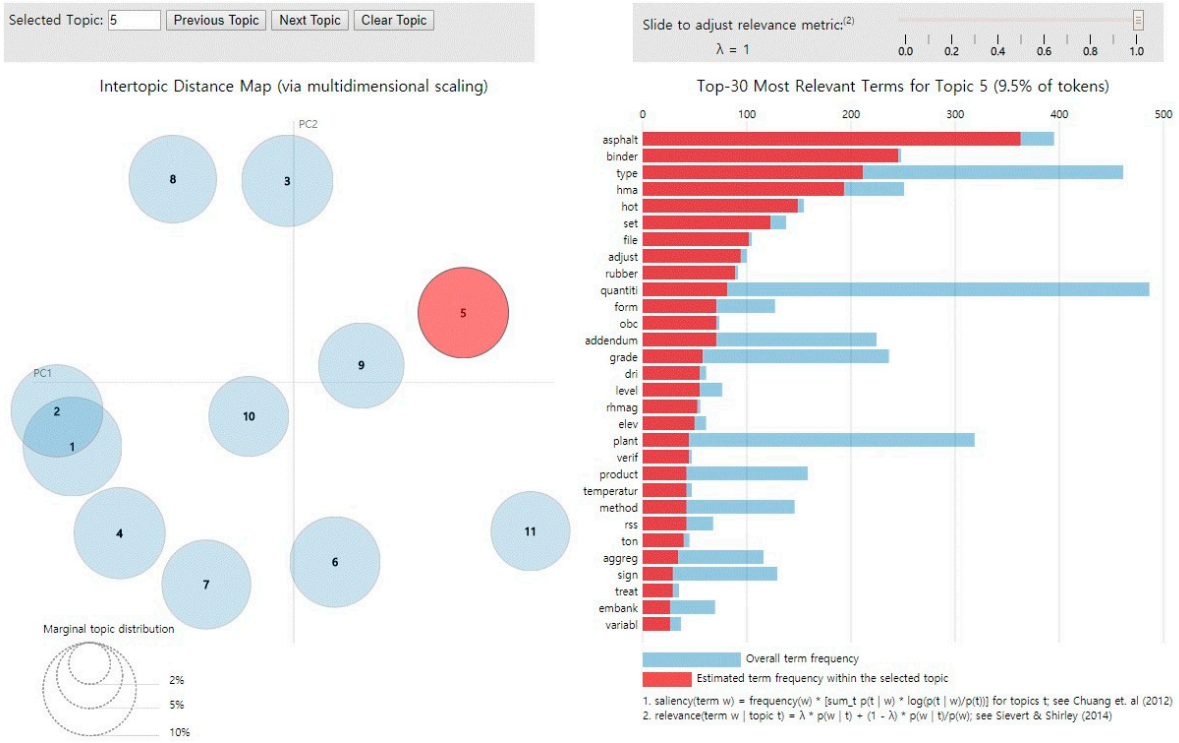


Figure 5. Topic Modeling Results using Latent Dirichlet Allocation (LDA).

4.2. Integrated Data Set

A data set for building the risk prediction model was created by integrating the value derived from unstructured text topic modeling and the initially collected numeric data. In addition to the text data, the 243 infrastructure projects ordered by Caltrans include numeric data, which consists of the number of bidders, the number of addendum issues, the number of Notice To Bidders (NTB) changes, the number of drawing changes, the number of specification changes, the number of bid item changes, the number of wage rate changes, and the pre-bid clarification of the project. The pre-bid clarification of the project was determined based on whether or not the project was clarified in advance. The unclarified status of each project was judged according to the following criteria:

- if the bidders have queried and the owner has not responded within a deadline;

- if the bidders have requested additional information on unclear information in the bidding document, but the owner has declined; and,
- if the bidders have made further proposals in case of danger and the owner has rejected.

When a total of 243 data were divided into three groups according to the bid risk percentage, and 89 low risk groups, 77 moderate risk groups, and 77 high risk groups were labeled. Based on the three classified groups, the bid risk prediction models learn the risk level of projects and classify each project into one of the three risk groups. In order to ensure the accuracy of the classification prediction results, it is preferable for the same number of data to be allocated for each group. In this study, the low risk group was assigned slightly more data than the other two groups, but because they did not show a large difference in the entire ratio, the data were analyzed for classification learning.

Table 2 shows part of the integrated data set. T01–T11 data are obtained from text topic modeling, and N01–N08 data is a numeric data set from bidding information. Each variable is defined as follows:

- N01 is the number of bidders; N02 indicates whether the project is pre-clarified or unclarified; N03 is the number of NTB changes; N04 is the number of drawings changes; N05 is the number of specification changes; N06 is the number of bid item changes; N07 is the number of wage rate changes; and, N08 is the number of issued addendums issued.

Table 2. Part of integrated data set.

Variables from Text Modeling											Variables from Numeric Data							
T01	**T02**	**T03**	**T04**	**T05**	**T06**	**T07**	**T08**	**T09**	**T10**	**T11**	**N01**	**N02**	**N03**	**N04**	**N05**	**N06**	**N07**	**N08**
0.083	0.108	0.096	0.096	0.096	0.083	0.146	0.070	0.070	0.070	0.083	4	1	0	0	0	0	1	1
0.083	0.083	0.083	0.068	0.113	0.128	0.083	0.068	0.157	0.068	0.068	5	1	0	1	1	0	1	1
0.044	0.063	0.217	0.044	0.073	0.188	0.111	0.073	0.063	0.063	0.063	11	1	0	1	1	1	1	2
0.056	0.056	0.068	0.081	0.105	0.056	0.056	0.068	0.266	0.093	0.093	7	1	1	1	0	0	0	1
0.143	0.110	0.069	0.078	0.078	0.208	0.061	0.086	0.069	0.061	0.037	7	1	2	1	1	3	1	6

Examining the variables in Table 2, the value ranges of variables from text modeling and variables from numeric data differ. If the range difference of independent variables is large, learning for classification prediction may not be performed properly. Therefore, by assigning the revised values to the variables from text modeling, most variables are distributed within a similar range, as shown in Figure 6.

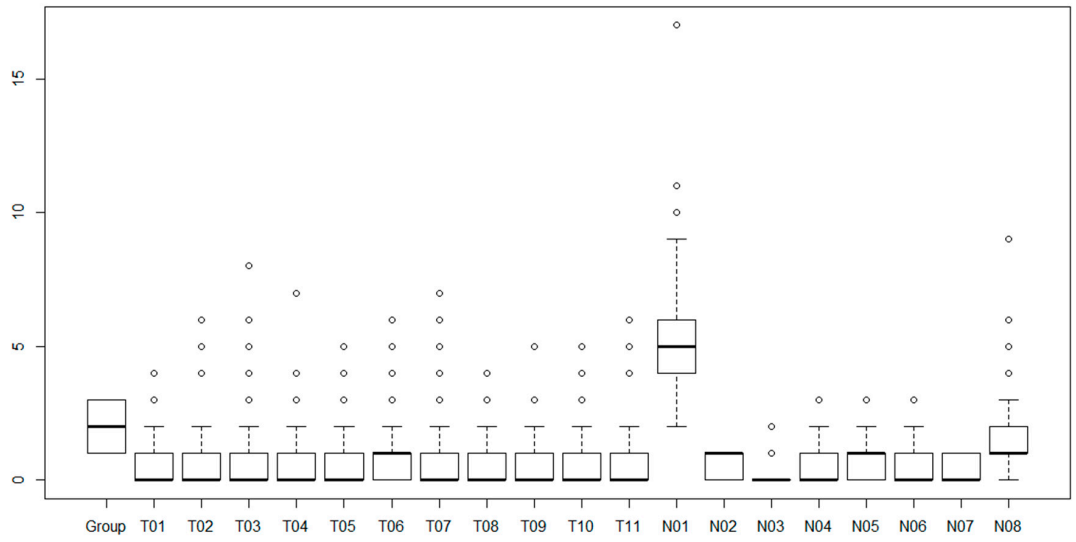


Figure 6. Integrated variables distribution.

4.3. Prediction Model Algorithm

Several different classification models were tested for bid risk prediction of the collected data. In this study, risk prediction modeling was performed using artificial neural networks (ANNs), support vector machines (SVM), k-nearest neighbors (KNN), and Naïve Bayes (NB) classification algorithms. The description of each classification model is as follows:

- ANNs: ANNs is a model that computes output values of artificial neurons by performing learning and assigning weights (connection strength of neurons) from a large number of input data. Input neurons are activated through sensors that sense the environment, and other neurons are activated through the weighted connections of previous active neurons [30]. One of the characteristics of ANNs is that they allow the development of sophisticated models through hidden layers.
- SVM: SVM is an algorithm for determining the optimal separating hyperplane that causes minimal generalization error [31]. SVM covers both linear and nonlinear classifiers [32]. For nonlinear classification, the given data needs to be mapped to the high dimensional feature space, and the kernel function is used to perform this efficiently.
- KNN: k-nearest neighbor classification is also called lazy learner. ANNs or SVMs first develop a model from the collected data and then apply it to the data to be classified, while KNN develops the model if data needs to be classified. The KNN algorithm classifies new objects according to the outcome of the closet object or the outcomes of several closest objects in the feature space of the training set [32]. Although one nearest data similar to the target data can be used for the classification, it is generally classified using k-nearest data.
- NB: The Naïve Bayes classifier considerably simplifies learning by assuming that the features are independent in each given class [33]. The Bayes classification is an algorithm that classifies groups with high posterior probability based on the knowledge of the prior probability and likelihood probability of each group using Bayes theorem.

5. Results

The prediction results using training data labeled as three risk groups (low risk group, moderate risk group, and high risk group) by four different classification algorithms are shown in Tables 3 and 4. Table 3 shows the results of risk prediction using only numeric data, while Table 3 shows the results of risk prediction using integrated data including text modeling results. The results show that the accuracy of the classification results is improved by about 20% using the integrated data compared to the numerical data only in all four classifiers. In addition, both the precision and recall results in Table 4 are considerably improved compared to Table 3.

Among the classification algorithms, SVM shows the highest classification performance in both Tables 3 and 4, and the classification results for the integrated data show that the accuracy of the classification model is as high as 72.92%. In addition, precision and recall were high for the high risk group, while the moderate group shows relatively low precision and recall. It is possible to determine the type of text topic that has been generated for each risk group from the results of classification modeling. Pre-bid RFI related to the type of bid item and quantity clarification issues (topic 7) were the most common in the low risk group, monitoring report, and duties (topic 9) related inquiries were most frequent in the moderate risk group, and the excavation quantity and payment (topic 4) related issued most frequently occurred in the high risk group (Table 5). In other words, uncertainty about the quantity and cost of rock excavation work was high in the high risk projects where the bidder's bidding price exceeded the owner's estimate by more than 10%. This implies that the bidding risk is high for projects in which uncertainty related to excavation quantity and payment has not been resolved in the infrastructure projects ordered by Caltrans.

Table 3. Results of classification with numeric data only.

Classification Algorithm	Accuracy	Class Precision			Class Recall		
		A	B	C	A	B	C
ANNs	45.83%	55.6%	41.7%	38.9%	55.6%	33.3%	46.7%
SVM	52.08%	57.9%	42.9%	53.3%	61.1%	40%	53.3%
KNN	50%	52.6%	46.2%	50%	55.6%	40%	53.3%
NB	45.83%	50%	333%	53.3%	50%	33.3%	53.3%

Note: A = low risk group; B = moderate risk group; C = high risk group.

Table 4. Results of classification with integrated data.

Classification Algorithm	Accuracy	Class Precision			Class Recall		
		A	B	C	A	B	C
ANNs	70.83%	63.6%	58.3%	92.9%	77.8%	46.7%	86.7%
SVM	72.92%	63.6%	63.6%	93.9%	77.8%	46.7%	93.3%
KNN	70.83%	64.7%	53.3%	93.8%	61.1%	53.3%	100%
NB	70.83%	71.4%	57.9%	86.7%	55.6%	73.3%	86.7%

Note: A = low risk group; B = moderate risk group; C = high risk group.

Table 5. Text topic according to each risk group.

	Text Topic	Key Terms
Low risk group (A)	Type of bid item and quantity clarification (Topic 7)	Type, quantity, depth, grade, location
Moderate risk group (B)	Monitoring report and duties (Topic 9)	Monitor, survey, permit, report, biologist
High risk group (C)	Excavation quantity and payment (Topic 4)	Excavation, remove, rock, backfill, pay

6. Discussion

In the case of predicting bid risks by using text data on pre-bid clarification, which represents the uncertainty of the bidding document, it is confirmed that prediction performance is greater than when predicted using numeric data only. In other words, the uncertain risk factors in the bidding document of each project obtained from the pre-bid RFI text data have a positive effect in increasing the accuracy of the risk prediction model. This means that unstructured text data contains useful information for predicting bid risks.

As a result, the information in the bidding document, which is an unstructured text data, is a major consideration when calculating the bidders' bid price. Furthermore, the results of this study enable inferences to be made about the type of uncertainty in the bidding document that increases the project's risk. That is, if the "quantity and cost of the rock excavation work" are not explicitly stated in the bid document, the bid risk can be increased. Moreover, when reviewing the bidding document for the infrastructure projects, uncertain information about the "rock excavation work" need to be clarified in advance through the pre-bid clarification process.

In addition, the precision and recall of the high risk group were higher than those of the other two groups. This means that the high risk group can be clearly distinguished from the other two groups. The reason why the high risk group has the highest prediction accuracy can be inferred from the text topic modeling results of each risk group. The topic of "excavation quantity and payment (topic 4)", which most frequently occurred in the pre-bid RFI of the projects belonging to the high risk group, was rarely occurred in the other two groups. However, the topic 9 and topic 7, which appeared in the moderate risk group and the low risk group, respectively, occurred in similar frequency in the other groups. As a result, the risk factors included in the high risk group projects were highly predictive because they are exclusive, unlike the other two groups.

In this study, risk prediction modeling based on uncertainty risk, which represents project characteristics, was conducted at the bidding stage of construction projects. However, various

influencing factors affect the bidder's bid price in the bidding decision process. Therefore, the risk predicted in this study may be only some of the factors that constitute the total bid risk. For this reason, some previous studies [13,14] showing the higher accuracy of constriction bidding risks than the accuracy of this study. In this study, however, the results of risk prediction modeling were improved by incorporating the unstructured text data of the bidding document, which was not considered by many previous studies. Moreover, the accuracy of the predicting results for the bid risks of the construction projects of this study is higher than the accuracy of the risk prediction model (44.47%) performed in [17], which predicted cost overruns through uncertainty risks on bidding document including text data.

7. Conclusions

In this study, construction bid risks were predicted by integrating with pre-bid clarification information, which determines the uncertainty of bid document, and numeric data since most of the important information that determines the uncertainty of a construction project is in the form of unstructured text data. As a result of the prediction modeling, the risk prediction accuracy was 72.92% by the SVM classifier using integrated data, while the prediction accuracy using numeric data only was 52.08% by the SVM classifier. This result proves the effectiveness of the proposed integrated data mining. In addition, this study presented the results of the types of uncertainty in the bidding document that increases a project's bidding risks. If the amount of analysis data increases and external factors, as well as project internal factors are taken into consideration, we expect more comprehensive bid risk predicting and decision models to be completed in future works.

Acknowledgments: This research was supported by a grant (16AUDP-B087012-03) from Urban Architecture Research Program funded by the Ministry of Land, Infrastructure and Transport of the Korean government.

Author Contributions: JeeHee Lee and June-Seong Yi conceived and designed the experiments; JeeHee Lee analyzed data and wrote the paper; June-Seong Yi provided technical support and revised the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ahmad, I.; Minkarah, I. Questionnaire survey on bidding in construction. *J. Manag. Eng.* **1988**, *4*, 229–243. [CrossRef]
2. Shash, A.A. Factors considered in tendering decisions by top UK contractors. *Constr. Manag. Econ.* **1993**, *11*, 111–118. [CrossRef]
3. Fayek, A. Competitive bidding strategy model and software system for bid preparation. *J. Constr. Eng. Manag.* **1998**, *124*, 1–10. [CrossRef]
4. Chua, D.K.H.; Li, D.Z.; Chan, W.T. Case-based reasoning approach in bid decision making. *J. Constr. Eng. Manag.* **2001**, *127*, 35–45. [CrossRef]
5. El-Mashaleh, M.S. Empirical framework for making the bid/no-bid decision. *J. Manag. Eng.* **2012**, *29*, 200–205. [CrossRef]
6. Brook, M. *Estimating and Tendering for Construction Work*, 3rd ed.; Butterworth-Heinemann: Oxford, UK, 2004.
7. Laryea, S. Quality of tender documents: Case studies from the UK. *Constr. Manag. Econ.* **2011**, *29*, 275–286. [CrossRef]
8. Duzkale, A.K.; Lucko, G. Exposing uncertainty in bid preparation of steel construction cost estimating: II. Comparative analysis and quantitative CIVIL classification. *J. Constr. Eng. Manag.* **2016**, *142*, 04016050. [CrossRef]
9. Lee, J.H.; Yi, J.S.; Son, J.W.; Jang, Y.E. Pre-bid clarification for construction project risk identification using unstructured text data analysis. In Proceedings of the Joint Conference on Computing in Construction, Heraklion, Greece, 4–7 July 2017; pp. 219–227.
10. Carr, P.G. Investigation of bid price competition measured through prebid project estimates, actual bid prices, and number of bidders. *J. Constr. Eng. Manag.* **2005**, *131*, 1165–1172. [CrossRef]

11. Zeng, J.; An, M.; Smith, N.J. Application of a fuzzy based decision making methodology to construction project risk assessment. *Int. J. Proj. Manag.* **2007**, *25*, 589–600. [CrossRef]
12. Leśniak, A.; Plebankiewicz, E. Modeling the decision-making process concerning participation in construction bidding. *J. Manag. Eng.* **2013**, *31*, 04014032. [CrossRef]
13. Han, S.H.; Diekmann, J.E. Approaches for making risk-based go/no-go decision for international projects. *J. Constr. Eng. Manag.* **2001**, *127*, 300–308. [CrossRef]
14. Han, S.H.; Diekmann, J.E. Making a risk-based bid decision for overseas construction projects. *Constr. Manag. Econ.* **2001**, *19*, 765–776. [CrossRef]
15. Mak, S.; Picken, D. Using risk analysis to determine construction project contingencies. *J. Constr. Eng. Manag.* **2000**, *126*, 130–136. [CrossRef]
16. Laryea, S.; Hughes, W. Risk and price in the bidding process of contractors. *J. Constr. Eng. Manag.* **2010**, *137*, 248–258. [CrossRef]
17. Williams, T.P.; Gong, J. Predicting construction cost overruns using text mining, numerical data and ensemble classifiers. *Autom. Constr.* **2014**, *43*, 23–29. [CrossRef]
18. Christodoulou, S. Optimum bid markup calculation using neurofuzzy systems and multidimensional risk analysis algorithm. *J. Comput. Civ. Eng.* **2004**, *18*, 322–330. [CrossRef]
19. Abotaleb, I.S.; El-Adaway, I.H. Construction Bidding Markup Estimation Using a Multistage Decision Theory Approach. *J. Constr. Eng. Manag.* **2016**, *143*, 04016079. [CrossRef]
20. Weiss, S.M.; Indurkhya, N.; Zhang, T. *Fundamentals of Predictive Text Mining*; Springer: London, UK, 2010; Volume 41.
21. Tan, A.H. Text mining: The state of the art and the challenges. In Proceedings of the PAKDD Workshop on Knowledge Discovery from Advanced Databases; Beijing, China, 26–28 April 1999; pp. 65–70.
22. Christopher, D.M.; Prabhakar, R.; Schütze, H. *Introduction to Information Retrieval*; Cambridge University Press: New York, NY, USA, 2008.
23. Steyvers, M.; Griffiths, T. *Probabilistic Topic Models. Handbook of Latent Semantic Analysis*; Routledge: New York, NY, USA, 2007; Volume 427, pp. 424–440.
24. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
25. Erosheva, E.A. Grade of Membership and Latent Structure Models with Application to Disability Survey Data. Ph.D. Thesis, Carnegie Mellon University, Pittsburgh, PA, USA, 2002.
26. Buntine, W.; Löfström, J.; Perkiö, J.; Perttu, S.; Poroshin, V.; Silander, T.; Tirri, H.; Tuominen, A.; Tuulos, V. A Scalable Topic-Based Open Source Search Engine. In Proceedings of the IEEE/WIC/ACM Conference on Web Intelligence, Beijing, China, 20–24 September 2004; pp. 228–234.
27. Griffiths, T.L.; Steyvers, M. *Finding Scientific Topics*; National Academy of Sciences: Washington, DC, USA, 2004; Volume 101, pp. 5228–5235.
28. Buntine, W.L. Estimating Likelihoods for Topic Models. In Proceedings of the 1st Asian Conference on Machine Learning, Nanjing, China, 2–4 November 2009; pp. 51–64.
29. Sievert, C.; Shirley, K.E. LDAvis: A method for visualizing and interpreting topics. In Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, Baltimore, MD, USA, 27 June 2014; pp. 63–70.
30. Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Netw.* **2015**, *61*, 85–117. [CrossRef] [PubMed]
31. Burges, C.J. Tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov.* **1998**, *2*, 121–167. [CrossRef]
32. Ledolter, J. *Data Mining and Business Analytics with R*; Wiley: Hoboken, NJ, USA, 2013.
33. Rish, I. An empirical study of the naive Bayes classifier. In Proceedings of the IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence, Seattle, WA, USA, 4–10 August 2001; Volume 3, pp. 41–46.