axioms

MDPI

Article

Text Data Analysis Using Generalized Linear Mixed Model and Bayesian Visualization

Sunghae Jun

Department of Big Data and Statistics, Cheongju University, Cheongju 28503, Republic of Korea; shjun@cju.ac.kr

Abstract: Many parts of big data, such as web documents, online posts, papers, patents, and articles, are in text form. So, the analysis of text data in the big data domain is an important task. Many methods based on statistics or machine learning algorithms have been studied for text data analysis. Most of them were analytical methods based on the generalized linear model (GLM). For the GLM, text data analysis is performed based on the assumption of the error included in the given data and follows the Gaussian distribution. However, the GLM has shown limitations in the analysis of text data, including data sparseness. This is because the preprocessed text data has a zero-inflated problem. To solve this problem, we proposed a text data analysis using the generalized linear mixed model (GLMM) and Bayesian visualization. Therefore, the objective of our study is to propose the use of GLMM to overcome the limitations of the conventional GLM in the analysis of text data with a zero-inflated problem. The GLMM uses various probability distributions as well as Gaussian for error terms and considers the difference between observations by clustering. We also use Bayesian visualization to find meaningful associations between keywords. Lastly, we carried out the analysis of text data searched from real domains and provided the analytical results to show the performance and validity of our proposed method.

Keywords: Bayesian inference and visualization; clustering; text data; generalized linear mixed model; big data analysis

MSC: 62C10; 62F15

Citation: Jun, S. Text Data Analysis Using Generalized Linear Mixed Model and Bayesian Visualization. *Axioms* **2022**, *11*, 674. https://doi.org/10.3390/axioms11120674

Academic Editor: Joao Paulo Carvalho

Received: 17 October 2022
Accepted: 25 November 2022
Published: 26 November 2022

1. Introduction

Since a large part of big data exists in the form of text, text data analysis has been an important topic studied in the big data areas. Most text data, such as web documents, papers, patents, and articles, are transformed into structured data for text analysis by statistics and machine learning algorithms [1–3]. A typical type of structured data is a matrix in which rows and columns are composed of documents and keywords, respectively [2,3]. This matrix is constructed through preprocessing using text-mining techniques [4]. Each element of this matrix represents the frequency of a keyword that occurred in a document. In general, since the number of keywords is very large compared to the number of documents, most elements in this matrix have a value of zero [5–7]. Because of this zero-inflated problem, the structured matrix data is very sparse. The sparsity decreases the performance of conventional analysis models based on the generalized linear model (GLM) [6]. The GLM uses the exponential family as output density, and the mean of output is a linear combination of inputs [8]. The linear combination is called a linear predictor in GLM. The GLM also uses the link function of output corresponding to the linear predictor of input [8]. We choose various functions, such as identity, logit, or log, for the link function [8]. By using the link functions, GLM can apply various suitable probability distributions to error terms [8]. Therefore, GLM was considered in various big data analysis tasks, including text data [3,9,10].

However, GLM shows limitations in analysis performance in the document keyword matrix, which has the problem of zero-inflated sparsity [6]. To overcome the sparseness of text data, we consider generalized linear mixed model (GLMM) and Bayesian visualization. The GLMM builds multi-task learning to apply the output to the group and item levels [8,11–15]. In this paper, the clustering approach improves the analytical performance of GLMM in text data analysis. Finally, we carry out Bayesian visualization using the results of GLMM. Therefore, the motivation of our research is to find a solution to the problem of the sparsity of zero-inflated problems that occurred in the process of text data analysis of patent documents. The rest of this paper is organized as follows. In Section 2, we describe the research background related to our research. Next, we show the proposed method in Section 3. The experiments and results are illustrated in Section 4, and the discussions are shown in Section 5. Finally, we show our conclusions and future works in the last section.

2. Research Background

In this section, we show two issues related to our proposed method. The first is the survey results of existing studies for text data analysis, and the second is the concept of the linear model (LM) of statistics.

2.1. Text Data Analysis

We can find interesting insights and hidden patterns in text documents using statistics and machine learning algorithms. Since a large part of big data consists of text, text data analysis is being treated as important in big data. We can consider two approaches to text data analysis. The first approach is text data analysis based on individual keywords. In this case, the aim of text data analysis is to find the statistical relations between the keywords [2,3,9,10]. Another approach is to use a topic that consists of several keyword combinations. That is, text data analysis is performed through model construction and visualization using topics [16–19]. In addition, text data analysis has been conducted in various ways and environments [20–23]. Figure 1 shows two approaches to text data analysis.

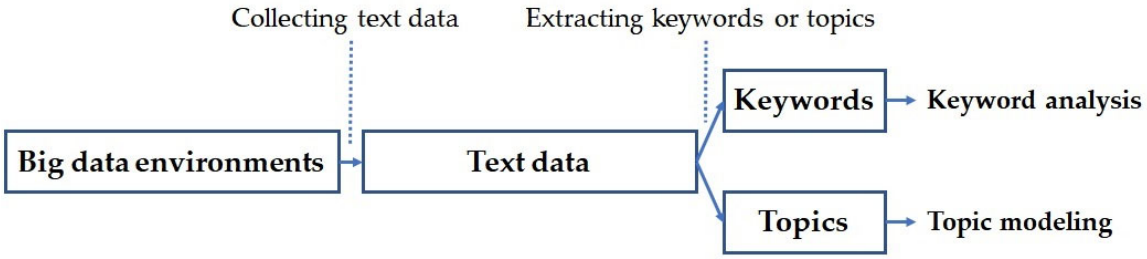

Figure 1. Two approaches to text data analysis.

In big data environments, there are huge amounts of data consisting of diverse data types, such as web pages, texts on social network services, sensing data on the internet of things, articles, papers, and patent documents. So, we can collect text data from various big data resources. From the collected text data, we extract keywords or topics for keyword analysis or topic modeling, respectively. Next, we apply the results by text data analysis to many applications.

2.2. Linear Model

In statistical data analysis, we try to build probability models that can best explain the dependent variable, y, using independent variables $X = (X_1, X_2, \ldots, X_p)$ [24,25]. In this case, the most basic is the linear model (LM) as shown in Equation (1) [26,27].

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon \quad (1)$$

where, $\beta = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p)$ is the parameter vector of LM. The error term, ε, is a random variable following a normal distribution with a mean of 0 and a variance of σ^2.

Additionally, the LM assumes that the distribution of y is normal. Using the given data with y and X, we estimate the parameters. In real problems, y follows various distributions, such as binomial and Poisson distributions as well as the normal distribution. However, the LM cannot support anything other than the normal distribution. To solve this limitation of LM, we consider GLM, which can use various probability distributions as well as normal distributions to explain y. For this flexible analysis, GLM has the following three characteristics. First, the linear predictor, η, is defined in Equation (2) [28].

$$\eta = X\beta \tag{2}$$

In Equation (2), X and β are independent variables and the model parameters, which were defined in Equation (1). That is, the linear predictor is calculated as the product of X and β and is used to predict y. The second component of GLM is the link function, $g(\cdot)$, defined by Equation (3) [8]

$$g(\mu_y) = X\beta \tag{3}$$

where, μ_y is the mean of y, and the link function is related to the linear predictor. In addition, the μ_y is represented by $g^{-1}(X\beta)$ in Equation (3). The last component of GLM is a random variable and probability distribution. Given X, μ_y is represented as $E(y|X)$. Using the linear predictor of ith observation, $X_i\beta$, we represent the distribution of y in Equation (4) [8].

$$p(y|X,\beta) = \prod_{i=1}^{n} p(y_i|X_i\beta) \tag{4}$$

In Equation (4), $p(\cdot)$ is one of the diverse probability distributions, such as the Poisson or binomial distributions, as well as the normal distribution. In addition, n is the number of data objects. So far, there have been many studies on data analysis using GLM [29–31]. The research results were used for data analysis tasks occurring in various fields such as medicine, bio, marketing, information technology, etc. However, in text data analysis, GLM has difficulties with some problems, such as differences between objects, sparsity, and grouping. This paper proposes an extension of GLM and Bayesian visualization to solve this problem.

3. Proposed Methods

Although most statistical analysis work has been performed based on GLM so far, in the text data field, it is difficult to use GLM as it is due to the difference between observed objects and data sparseness by zero-inflated problem [5–7]. So, in this paper, we propose a method of text data analysis to overcome the limitation of GLM in text data analysis. Our proposed method consists of four procedures, from the collection of text data to the visualization of analytical results.

3.1. Preprocessing of Documents

For text data analysis, we first need to perform preprocessing on the collected text documents. This is because most statistical methods and machine learning algorithms require structured data preprocessed by text-mining techniques. Text mining has various functions necessary for preprocessing text data. Figure 2 illustrates a text-mining process from text documents to the document–keyword matrix.

The collected text documents are transformed into a text corpus by natural language processing. Corpus is a text document collection. The text databases are built from the text corpus using parsing. Finally, we extract keywords from text databases and construct a document–keyword matrix. This is structured data for statistical analysis and machine learning.

3.2. Structure of Text Data

Most text data exist in the form of documents such as web pages, articles, papers, reports, or patents. We have to collect the documents related to the target for text data analysis. In this paper, we consider patent documents as text data. A patent contains

various information on research and development technology, such as the title, abstract, inventor, applicant, claims, drawings, citation information, and technology classification code [32,33]. So, we used patent data to perform technical analysis for the target domain. To collect patent text data, we searched patent documents from patent databases around the world [34,35]. We make the keyword search expression related to the target technology for patent retrieval and transform the collected documents into structured data for statistical data analysis. For text data analysis, statistical methods and machine learning algorithms require structured data types, such as relational database tables [7,10]. The table consists of rows and columns representing patent documents and patent keywords, respectively. In this paper, the collected patent documents are represented in Equation (5).

$$PDs = (pd_1,\ pd_2,\ \ldots,\ pd_n) \tag{5}$$

In Equation (5), the patent documents (*PDs*) consist of n patent documents. Next, we will preprocess the *PDs* to build structured data for cluster analysis, GLMM, and Bayesian visualization in sequence. We will use text-mining techniques for data preparation, import, cleaning, and removing stopwords for constructing structured patent data [4] to obtain the patent keyword matrix as structured data. This matrix consists of patent documents and keywords for rows and columns, respectively. The element of this matrix is the frequency of a keyword included in each patent document. Table 1 shows the patent keyword matrix of our research.

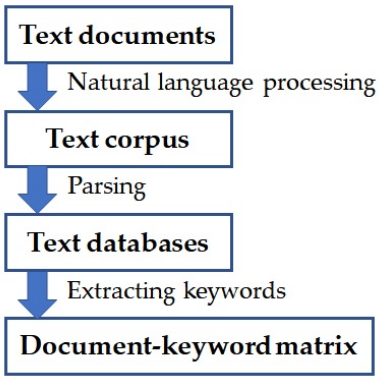

Figure 2. Text-mining process.

Table 1. Patent keyword matrix.

Patent Document	$Keyword_1$	$Keyword_2$	$\cdots$	$Keyword_p$
pd_1	$Freq_{11}$	$Freq_{12}$	$\cdots$	$Freq_{1p}$
pd_2	$Freq_{21}$	$Freq_{22}$	$\cdots$	$Freq_{2p}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
pd_i	$Freq_{i1}$	$Freq_{i2}$	$\cdots$	$Freq_{ip}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
pd_n	$Freq_{n1}$	$Freq_{n2}$	$\cdots$	$Freq_{np}$

The dimension of this matrix is $(n \times p)$ and its element $Freq_{ij}$ represents the occurred frequency of $Keyword_j$ in ith patent document (pd_i). In the proposed method, the keyword and patent document are used as variables and observations, respectively.

3.3. Clustering

In this paper, we perform the GLMM to determine the significant keywords (variables) for Bayesian visualization. Since the GLMM requires the number of clusters for the given data, we need to find the optimal number of clusters for the patent keyword matrix [15]. So, we consider Silhouette analysis as a method to determine the number of clusters optimally.

The Silhouette score is the average of silhouette coefficients of all patent documents [36–38]. The Silhouette coefficient (SC) of a patent document is calculated as Equation (6) [36,38].

$$\mathrm{SC} = \frac{(d_2 - d_1)}{max(d_1, d_2)} \tag{6}$$

In Equation (6), d_1 represents the average distance to other documents in the same cluster. d_2 is the average distance from documents belonging to the nearest cluster. The SC has a value from -1 to 1, and the closer to 1, the better the clustering, and the closer to -1, the poorer the clustering. Therefore, we determined the number of clusters with the largest SC value as the optimal number of clusters. Using the clustering result, we carried out GLMM in the next section.

3.4. Generalized Linear Mixed Model for Text Data Analysis

Although GLM is superior to the LM because it can be modeled even as a nonlinear function, it has difficulties in the mixed model for clustered and correlated data [13,15]. Additionally, the GLM has a limitation in analyzing the sparse data with the zero-inflated problem [6,7]. In contrast, the GLMM provides fixed and random effects for mixed models as well as clustered and correlated models [13,15]. In addition, GLMM has been used to analyze the data with overdispersion or nonlinear [8,11,12]. So, we consider GLMM to analyze text data with the zero-inflated problem. The preprocessed text data from text documents is a matrix with patent documents and keywords. This is structured data for data analysis based on statistics and machine learning algorithms. The element of this matrix indicates the frequency of a keyword appearing in each patent document. In the GLMM, we build a multi-task model for the response to contain the information on the patent group and patent document [8]. The GLMM involves fixed and random effects similar to the generalized mixed models, but the probability distribution of error can be any probability distribution belonging to the exponential family [11]. The GLMM is defined as the following in Equation (7) [11].

$$g(\mu) = X\beta + Wv \tag{7}$$

μ is the expectation of y given v, $E(y|v)$, where y and v are the response vector and vector of the random model effect, respectively. X is the data matrix. In addition, β and W are the vectors of fixed effects and constant matrix related to random effects. In Equation (7), the model parameters are estimated by solving mixed model equations based on the generalized form [12]. The patent documents related to target technology consist of various sub-technologies. For example, the patents associated with artificial intelligence consist of many detailed technologies such as machine learning, sensors, big data, communication, robotic automation, machine vision, optimization, soft system development, etc. So, we consider the sub-technology clusters in patent analysis. Figure 3 shows the patent document data and the corresponding GLMM with clusters.

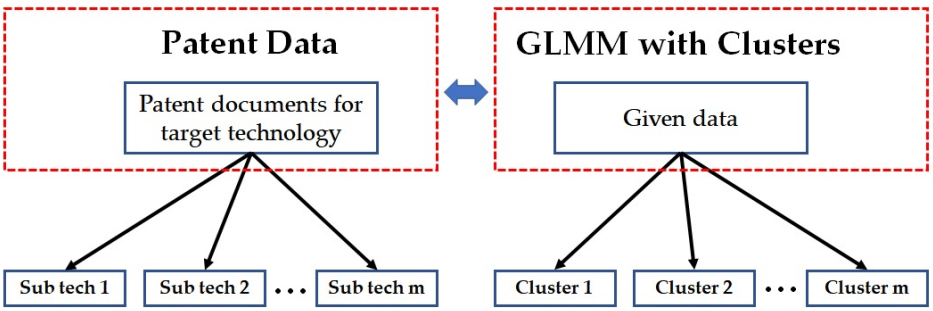

Figure 3. Patent data and GLMM with clusters.

Figure 3 describes the patent documents composed of m sub-technologies (Sub tech 1, Sub tech 2, ... , Sub tech m) and the GLMM with data having m clusters (Cluster 1, Cluster 2, ... , Cluster m) corresponding to the patent data. Using the GLMM with clusters, we analyze the patent keyword matrix to understand the target technology. In general, we have to partition the patent data into many sub-technology clusters for efficient patent analysis. However, in previous studies related to patent analysis, there was no study on patent analysis methods considering sub-technology clusters. In this paper, we study a text analysis method considering sub-technologies on target technology. The data form of our model is shown in Equation (8).

$$(x_{i1},\ x_{i2},\ \ldots,\ x_{ip},\ y_i), i = 1, 2, \ldots, n \tag{8}$$

$(x_1,\ x_2,\ \ldots,\ x_p)$ are p explanatory variables (patent keywords) and y is a response variable. n is the number of patent documents. The GLMM of patent data is represented in Equation (9) [8],

$$y = X\beta + Wv + \varepsilon \tag{9}$$

where $v \sim N(0,\ \Sigma_v)$ and $\varepsilon \sim N(0,\ \Sigma_\varepsilon)$. In addition, v and ε are independent of each other. So, we obtained the probability distribution of y from Equation (10) [8].

$$y \sim N(X\beta, W\Sigma_v W' + \Sigma_\varepsilon) \tag{10}$$

Next, we estimated the model parameters by the marginal likelihood function with Equation (11) [8],

$$L(\beta, \Sigma|y) = \int \prod_{i=1}^{n} f(y_i|v, \beta, \phi_i) f(v|\Sigma) dv \tag{11}$$

where ϕ_i is the known weight. We maximized this function to obtain the estimates of model parameters. To evaluate model performance, we used Akaike's information criteria (AIC) in Equation (12) [25].

$$AIC = 2p + nlog\left(\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}\right) \tag{12}$$

$\hat{y}$ is the predicted value of y. In addition, we checked the statistical significance of the explanatory variables for the response variables through hypothesis testing on the coefficients of each variable in the model in Equation (13) [27].

$$H_0 : \beta = 0 \ \ versus \ \ H_1 : \beta \neq 0 \tag{13}$$

We used the p-value (probability value) to perform the hypothesis testing in Equation (13). For example, if the p-value is less than 0.05, then we reject the H_0 under 95% confidence level. This means that the explanatory variable is statistically significant to the response variable [27]. Therefore, we selected the explanatory variables with a p-value less than 0.05 and used them for Bayesian visualization in the next section.

3.5. Bayesian Visualization

For Bayesian visualization, we defined a network structure, $G = (S,\ t)$, in which S and t represent node and connection, respectively [39,40]. The main goal of Bayesian visualization is to express the technology relationships among the patent keywords. In this paper, the node is defined in Equation (14),

$$S = \{Keyword_1, keywoord_2, \ldots, Keyword_s\} \tag{14}$$

where s is the number of the keywords selected from the results of GLMM. t_{ij} is defined as a connection between $Keyword_i$ and $Keyword_j$. We built a directed acyclic graph (DAG) containing only directed connections using the following chain rule in Equation (15) [39,40].

$$f(X) = \prod_{i=1}^{s} f(X_i|\Pi_{X_i}) \quad (15)$$

X_i represents $Keyword_i$ and Π_{X_i} are the parent nodes of X_i. Additionally, each keyword, X_i, is conditionally independent of other non-descendants. So, we carried out Bayesian visualization using DAG networks based on Equation (15). Figure 4 shows the procedures of our proposed methods.

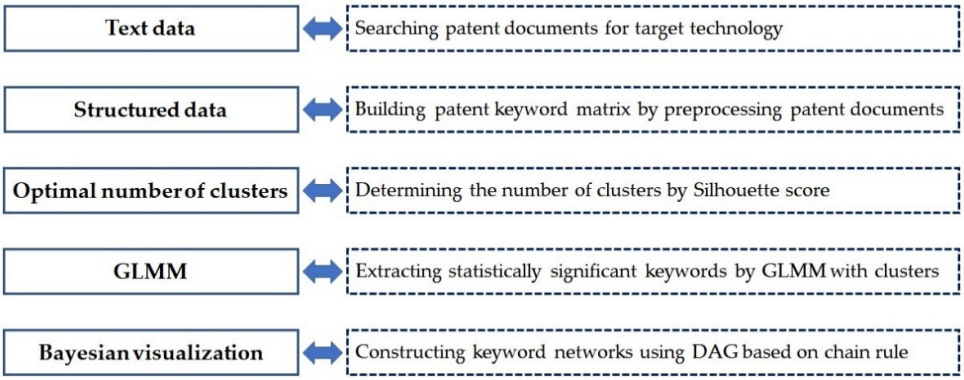

Figure 4. Procedures of the proposed methods.

First, we searched the patent documents from patent databases for text data analysis. Next, we built a patent keyword matrix as structured data by preprocessing patent documents. For performing the GLMM, we have to determine the optimal number of clusters. So, we calculated the Silhouette scores to choose the number of clusters. We extracted the statistically significant keywords from the GLMM results and used them to construct the keyword networks by DAG based on the chain rule. Finally, we applied the visualization results to various domain tasks, such as technology analysis and forecasting. In the next section, we show our experimental results to illustrate how our proposed methods can be applied to practical tasks.

4. Experiments and Results

To illustrate the performance and validity of our proposed methods, we searched the patent documents related to the technology of intelligent disaster prevention from the patent database [34,35]. A total of 16,875 patent documents were selected through the valid patent selection process. In addition, 100 keywords were selected using text-mining preprocessing [4]. For our experiments, we determined the top 20 keywords based on their occurred frequency values. Table 2 shows the total and top 20 keywords extracted from the patent documents.

In the top 20 keywords of Table 2, the frequency of each keyword is shown in parentheses after a keyword. We used the top 20 keywords to evaluate the performance of the proposed methods. First, we performed the cluster analysis by Silhouette score. Table 3 shows the clustering results of the patent data through the Silhouette coefficient.

In the clustering, we calculated the Silhouette coefficient values in order to select the optimal number of clusters between 3 and 9. From the results in Table 3, we determined 3, which has the largest Silhouette coefficient value, as the optimal number of clusters. Using the clustering result, we carried out GLMM and compared the result of GLMM with GLM. Table 4 shows the comparison of performance between GLMM and GLM.

Table 2. Patent keyword lists.

Keyword	List
Total keywords	Abnormal, acoustic, air, alarm, analysis, artificial, audio, automatic, battery, big, camera, cloud, cluster, communication, computing, damage, data, database, deep, detection, device, digital, earth, earthquake, electric, energy, engine, engineering, environment, estimation, fault, feedback, fire, flow, fluid, forecast, fuzzy, gas, geological, grid, health, human, image, information, intelligence, interaction, interface, land, laser, learning, light, machine, magnetic, map, measurement, memory, mobile, monitoring, network, neural, normal, oil, optical, pattern, picture, power, pressure, probability, protocol, pulse, radar, radio, remote, risk, robot, rock, sampling, satellite, sea, security, sensor, signal, software, space, spatial, speed, stability, statistics, temperature, time, underground, vehicle, velocity, video, warning, water, weather, web, wind, wireless
Top 20 keywords	Data (117,091), analysis (39,875), information (28,972), time (27,115), signal (26,292), device (25,149), power (19,739), image (19,602), network (17,646), monitoring (16,454), fault (15,723), detection (13,475), sensor (11,862), temperature (11,208), environment (10,172), machine (9744), water (8807), wind (8311), cloud (8020), communication (7679)

Table 3. Silhouette coefficient for determining the number of clusters.

Clustering	Number of Clusters						
	3	**4**	**5**	**6**	**7**	**8**	**9**
Average Silhouette coefficient	0.21	0.11	0.12	0.11	0.08	0.11	0.09

Table 4. Comparison of performance between GLMM and GLM.

Parameter and AIC	Significance Probability	
	GLMM	**GLM**
data	0.0001	0.0036
analysis	0.0001	0.0086
information	0.0077	0.0034
time	0.6540	0.0862
signal	0.0001	0.0001
device	0.1250	0.0060
power	0.0448	0.0904
image	0.0009	0.0105
network	0.0001	0.1000
monitoring	0.0001	0.0001
fault	0.0001	0.0001
detection	0.0049	0.1004
sensor	0.0001	0.0001
temperature	0.0001	0.0001
environment	0.0194	0.4469
machine	0.1100	0.0135
water	0.9860	0.3872
cloud	0.6580	0.2273
communication	0.0582	0.0946
AIC	8603	85,487

In Table 4, the dependent variable is water, and all other variables are used as independent variables. We also represent the p-values of model parameters and the AIC value of each model. Since the AIC value of GLMM is much smaller than that of GLM, it can be confirmed that the performance of GLMM is better than GLM. In addition, the parameters with a p-value of less than 0.05 (95% confidence level) are statistically significant to explain the response variable (keyword) wind. We found that the number of significant parameters of GLMM is larger than GLM. Except for time, signal, machine, wind, cloud, and communication, all independent variables explain wind as statistically significant in

GLMM. Next, we show the Bayesian visualization by a construct learning algorithm based on the score in Figure 5.

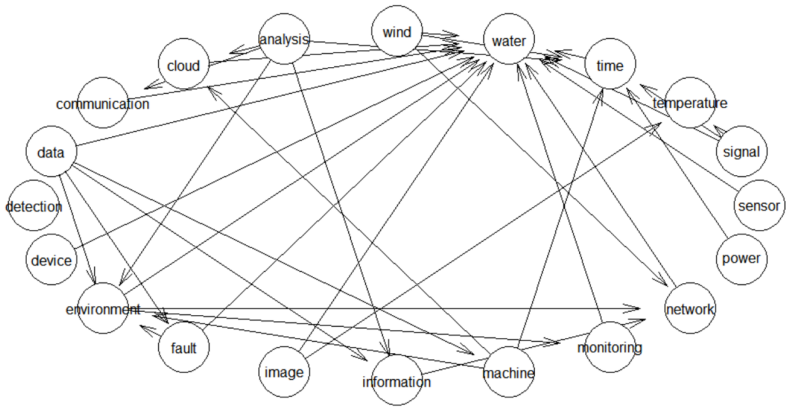

Figure 5. Bayesian visualization by a construct learning algorithm based on score.

We could confirm that most of the keywords are affecting the keyword water. The keyword data only affects other keywords such as water, machine, information, fault, or environment. So, we can conclude the technologies based on water, machine, information, fault, or environment are influenced by data technology. For example, the technology of fault is dependent on the technology of data, such as surveys, monitoring, management, and analysis. Moreover, data technology does not directly affect water technology. Instead, data technology influences information technology and indirectly affects water technology by enabling information technology to influence water technology. Using the directed acyclic graph, we show another approach to Bayesian visualization in Figure 6.

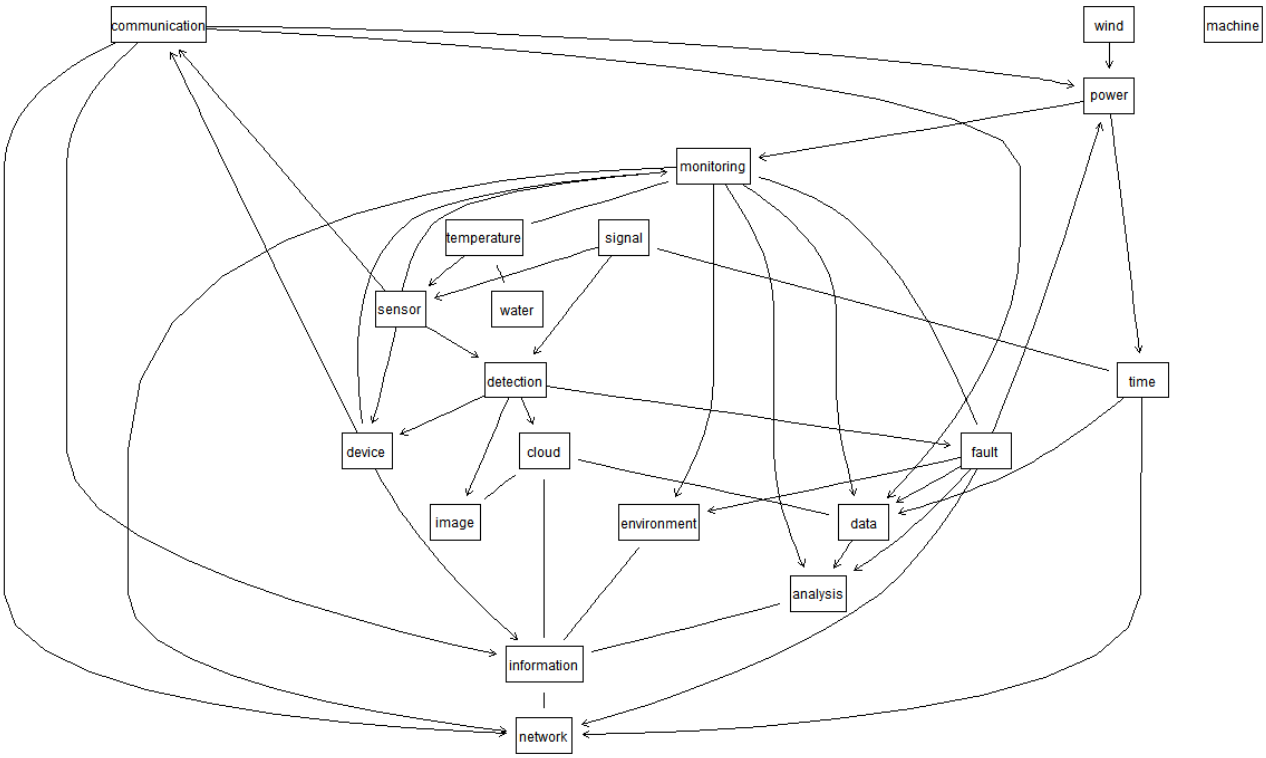

Figure 6. Bayesian visualization by the directed acyclic graph.

In Figure 6, we found that the keyword water is connected to the keyword temperature, and the temperature is also connected to the keyword monitoring. The keywords of communication, fault, time, and monitoring affect the keyword data, and the data directly

affects the keyword analysis. The analysis is connected to water through information, cloud, detection, sensor, and temperature. Finally, most keywords are connected to the keywords of information and network. Therefore, we concluded that the technology of information networks is the most basic technology in the field of intelligent disaster prevention.

5. Discussion

The zero-inflated problem occurs in the process of analyzing text data using statistics and machine learning and is the cause of the decreasing performance of the constructed model. We tried to overcome this problem with a statistical approach called the GLMM and showed the improved performance of our text data analysis by experimental results using practical patent documents. Our research will contribute to various fields of text data analysis, such as biomedicine, customer marketing, and social network service, as well as patent technology. In this paper, we used the AIC metric to evaluate the performance of the proposed method and the conventional method. In the future, more metrics for model evaluation will be developed for the efficient evaluation of newly developed models.

6. Conclusions

In this paper, we proposed a method for text data analysis. We used the GLMM to solve the sparsity of zero-inflated problems in patent keyword data. The GLMM with clustered data showed better performance than the traditional GLM. We also used the Silhouette score to determine the number of clusters optimally. This result was applied to the GLMM. Furthermore, we perform Bayesian visualization using the results of GLMM to find the relations between keywords.

We considered patent data analysis as one of the text data analyses. We collected the patent documents related to intelligent disaster prevention for our experiments. Using text-mining techniques, we build a patent keyword matrix as structured data for statistical analysis. We analyzed the matrix by conventional GLM and our GLMM for performance comparison between comparative models. From the experimental results, we found that the performance of GLMM is better than GLM. Using the GLMM results, we carried out Bayesian visualization to construct the networks of patent keywords for understanding the technological relations between sub-technologies of intelligent disaster prevention.

We applied the GLMM results to Bayesian networks to find the technology structure of the target technology. However, there are so many methods for big data visualization, such as social networks. In our future works, we will study more advanced methods for keyword visualizations from the results of GLMM.

Funding: This research received no external funding.

Data Availability Statement: Data is contained within the article.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Choi, S.; Park, S.; Jun, S. Text Data Analysis using Bayesian Quantile Regression and Multidimensional Scaling. *J. Korean Inst. Intell. Syst.* **2021**, *31*, 177–183.
2. Park, S.; Jun, S. Technological cognitive diagnosis model for patent keyword analysis. *ICT Express* **2020**, *6*, 57–61. [CrossRef]
3. Park, S.; Jun, S. Patent Keyword Analysis of Disaster Artificial Intelligence Using Bayesian Network Modeling and Factor Analysis. *Sustainability* **2020**, *12*, 505. [CrossRef]
4. Feinerer, I.; Hornik, K. *Package 'tm' Version 0.7-8, Text Mining Package*; CRAN of R Project, R Foundation for Statistical Computing: Vienna, Austria, 2022.
5. Jun, S.; Park, S.; Jang, D. Document Clustering Method Using Dimension Reduction and Support Vector Clustering to Overcome Sparseness. *Expert Syst. Appl.* **2014**, *41*, 3204–3212. [CrossRef]
6. Uhm, D.; Jun, S. Zero-Inflated Patent Data Analysis Using Generating Synthetic Samples. *Future Internet* **2022**, *14*, 211. [CrossRef]
7. Kim, J.M.; Jun, S. Zero-inflated Poisson and negative binomial regressions for technology analysis. *Int. J. Softw. Eng. Its Appl.* **2016**, *10*, 431–448. [CrossRef]
8. Murphy, K.P. *Machine Learning: A Probabilistic Perspective*; MIT Press: Cambridge, MA, USA, 2012.

9. Kim, J.; Jun, S. Graphical Causal Inference and Copula Regression Model for Apple Keywords by Text Mining. *Adv. Eng. Inform.* **2015**, *29*, 918–929. [CrossRef]
10. Park, S.; Jun, S. Patent Analysis Using Bayesian Data Analysis and Network Modeling. *Appl. Sci.* **2022**, *12*, 1423. [CrossRef]
11. Stroup, W.W. *Generalized Linear Mixed Models: Modern Concepts, Methods and Applications*; CRC press: Boca Raton, FL, USA, 2012.
12. Berridge, D.M.; Crouchley, R. *Multivariate Generalized Linear Mixed Models Using R*; CRC press: Boca Raton, FL, USA, 2012.
13. Mizdrak, P. Clustering Profiles in Generalized Linear Mixed Models Settings Using Bayesian Nonparametric Statistics. Ph.D. Thesis, Carleton University, Ottawa, ON, Canada, 2018.
14. Lee, J. A Study for Recent Development of Generalized Linear Mixed Model. *Korean J. Appl. Stat.* **2000**, *13*, 541–562.
15. Broström, G.; Jin, J.; Holmberg, H. *Package 'glmmML' Ver. 1.1.3, Generalized Linear Models with Clustering*; CRAN of R Project, R Foundation for Statistical Computing: Vienna, Austria, 2022.
16. Di Corso, E.; Proto, S.; Vacchetti, B.; Bethaz, P.; Cerquitelli, T. Simplifying Text Mining Activities: Scalable and Self-Tuning Methodology for Topic Detection and Characterization. *Appl. Sci.* **2022**, *12*, 5125. [CrossRef]
17. Allan, J.; Carbonell, J.G.; Doddington, G.; Yamron, J.; Yang, Y. Topic detection and tracking pilot study. In Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, Lansdowne, VA, USA, 8–11 February 1998; pp. 1–25.
18. Nakov, P.; Popova, A.; Mateev, P. Weight functions impact on LSA performance. In Proceedings of the Euro Conference RANLP, online, 1–3 September 2021; pp. 1–7.
19. Corso, E.D.; Proto, S.; Cerquitelli, T.; Chiusano, S. Towards automated visualisation of scientific literature. In Proceedings of the European Conference on Advances in Databases and Information Systems, Bled, Slovenia, 8–11 September 2019; pp. 28–36.
20. Saxena, G.; Santurkar, S. An iterative MapReduce framework for sports-based tweet clustering. In Proceedings of the Sixth International Conference on Computer and Communication Technology, Allahabad, India, 25–27 September 2015; pp. 9–14.
21. Bouaziz, A.; Pereira, C.C.; Pallez, C.D.; Precioso, F. Interactive generic learning method (IGLM): A new approach to interactive short text classification. In Proceedings of the 31st Annual ACM Symposium on Applied Computing, Pisa, Italy, 4–8 April 2016; pp. 847–852.
22. Duchrow, T.; Shtatland, T.; Guettler, D.; Pivovarov, M.; Kramer, S.; Weissleder, R. Enhancing navigation in biomedical databases by community voting and database-driven text classification. *BMC Bioinform.* **2009**, *10*, 317. [CrossRef] [PubMed]
23. Gross, O.; Doucet, A.; Toivonen, H. Language-independent multi-document text summarization with document-specific word associations. In Proceedings of the 31st Annual ACM Symposium on Applied Computing, Pisa, Italy, 4–8 April 2016; pp. 853–860.
24. Hogg, R.V.; Tanis, E.A.; Zimmerman, D.L. *Probability and Statistical Inference*, 9th ed.; Pearson: Essex, UK, 2015.
25. Bruce, P.; Bruce, A.; Gedeck, P. *Practical Statistics for Data Scientists*, 2nd ed.; O'Reilly Media: Sebastopol, CA, USA, 2020.
26. Hogg, R.V.; Mckean, J.W.; Craig, A.T. *Introduction to Mathematical Statistics*, 8th ed.; Pearson: Essex, UK, 2020.
27. Ross, S.M. *Introduction to Probability and Statistics for Engineers and Scientists*, 4th ed.; Elsevier: Seoul, Republic of Korea, 2012.
28. Gelman, A.; Carlin, J.B.; Stern, H.S.; Dunson, D.B.; Vehtari, A.; Rubin, D.B. *Bayesian Data Analysis*, 3rd ed.; Chapman & Hall/CRC Press: Boca Raton, FL, USA, 2014.
29. Sun, Y.; Wang, Q. An adaptive group LASSO approach for domain selection in functional generalized linear models. *J. Stat. Plan. Inference* **2022**, *219*, 13–32. [CrossRef]
30. Park, J.; Kang, S. Hierarchical Generalized Linear Models for Multiregional Clinical Trials. *Stat. Biopharm. Res.* **2022**, *14*, 358–367. [CrossRef]
31. Adesina, O.; Agunbiade, D.; Oguntunde, P. Flexible Bayesian Dirichlet mixtures of generalized linear mixed models for count data. *Sci. Afr.* **2021**, *13*, e00963. [CrossRef]
32. Hunt, D.; Nguyen, L.; Rodgers, M. *Patent Searching Tools & Techniques*; Wiley: Hoboken, NJ, USA, 2007.
33. Roper, A.T.; Cunningham, S.W.; Porter, A.L.; Mason, T.W.; Rossini, F.A.; Banks, J. *Forecasting and Management of Technology*; John Wiley & Sons: Hoboken, NJ, USA, 2011.
34. KIPRIS. Korea Intellectual Property Rights Information Service. Available online: www.kipris.or.kr (accessed on 1 July 2022).
35. USPTO. The United States Patent and Trademark Office. Available online: http://www.uspto.gov (accessed on 1 July 2022).
36. Batool, F.; Hennig, C. Clustering with the Average Silhouette Width. *Comput. Stat. Data Anal.* **2021**, *158*, 107190. [CrossRef]
37. Lovmar, L.; Ahlford, A.; Jonsson, M.; Syvanen, A.C. Silhouette scores for assessment of SNP genotype clusters. *BMC Genom.* **2005**, *6*, 35. [CrossRef] [PubMed]
38. Han, J.; Kamber, M.; Pei, J. *Data Mining: Concepts and Techniques*, 3rd ed.; Morgan Kaufmann: Waltham, MA, USA, 2012.
39. Scutari, M.; Denis, J.B. *Bayesian Networks with Examples in R*, 2nd ed.; Chapman & Hall/CRC: Boca Raton, FL, USA, 2021.
40. Nagarajan, R.; Scutari, M.; Lebre, S. *Bayesian Networks in R with Application and System Biology*; Springer: London, UK, 2013.