applied sciences

MDPI

Article

Zero-Inflated Patent Data Analysis Using Compound Poisson Models

Sangsung Park and Sunghae Jun *

Department of Statistics, Cheongju University, Cheongju 28503, Republic of Korea
* Correspondence: shjun@cju.ac.kr; Tel.: +82-10-7745-5677; Fax: +82-43-229-8432

Abstract: A large part of big data consists of text documents such as papers, patents or articles. To analyze text data, we have to preprocess the text documents and build a structured data based on a document-word matrix using various text mining techniques. This is because statistics and machine learning algorithms used in text analysis require structured train data. The row and column of the matrix are document and word, respectively. The element of the matrix represents the frequency value of the word occurring in each document. In general, because the number of words is much larger than the number of documents, most elements have zero values. Due to the sparsity problem caused by inflated zeros, the performance of the predictive model has decreased. In this paper, we propose a method to solve the sparsity problem and improve the model performance in text data analysis. We perform compound Poisson linear modeling to make the proposed method. To show the performance of our proposed method, we collect and analyze the patent documents from patent databases. In our experimental results, we compared the value of the Akaike information criterion (AIC) of the proposed model with traditional models, such as linear model, generalized linear model and zero-inflated Poisson model. Additionally, we illustrated that the AIC value of our proposed model is smaller than others. Therefore, we verify the validity of this paper.

Keywords: zero-inflated data; compound Poisson model; generalized linear model; Poisson distribution; document-word matrix

Citation: Park, S.; Jun, S. Zero-Inflated Patent Data Analysis Using Compound Poisson Models. *Appl. Sci.* **2023**, *13*, 4505. https://doi.org/10.3390/app13074505

Academic Editor: Zhi-Ting Ye

Received: 3 February 2023
Revised: 23 February 2023
Accepted: 31 March 2023
Published: 2 April 2023

1. Introduction

Since most of the information is in the form of text documents, the proportion of text data in the big data environment continues to increase [1–3]. Therefore, many studies for text data analysis have been conducted in various fields [2–4]. Because the analysis methods of statistics and machine learning cannot analyze text documents as they are, we must preprocess the collected text documents to enable analysis. In the preprocessing of text documents, we use text mining techniques and transform the documents into document-word matrix [5,6]. The row and column of this matrix represent document and word, respectively. Its elements are frequency values of words occurred in a document. In general, this matrix has a very sparse structure because a word occurring even once among all documents constitutes one column in the matrix [7,8]. That is, the most values of the elements are zeros because the number of words is much larger than the number of documents [7,8]. This is the zero-inflated problem [9–12]. Due to this problem, the document-word matrix has a very sparse data structure. Because of the sparsity, it is difficult to analyze the text document data using convenient methods such as the linear regression model. Research related to zero-inflated text data analysis were carried out in statistics and machine-learning areas [7–11,13–16]. Most of them used the probability distributions such as Poisson or negative binomial [9,11]. However, most of them had a limitation of model performance because of the sparsity of zero inflation.

Therefore, the motivation of this paper is to solve the zero-inflated sparsity problem in the structured text data. First, we survey the theoretical structure of traditional models

for zero-inflated data analysis and understand the limitations of the traditional models. Second, we apply the compound Poisson distribution to construct the analytical model for zero-inflated data. In addition, we will show the validity of the proposed method by carrying out performance experiments using real text documents to compare the new proposed and existing analysis models. The objective of our proposed method is to research and develop a predictive model using compound Poisson distribution for an efficient analysis of sparse text data. This is a mixture model of degenerate distribution for zero and continuous distribution for positive numbers greater than or equal to one [17]. The contribution of this paper is to improve the explanatory power of the predictive model by solving the problem of zero excess of text data that occurs in various big data fields. Our research can also expand the contribution of this paper by resolving the zero-inflated problem that occurs in various environments such as the sensor data of the internet of things (IoT), various learning data for artificial intelligence (AI), and medical image video data. For instance, the observed values from the sensing system of the IoT environment are mostly zero values. This is because most of the results of the sensing system of IoT do not change unless a special event occurs. Consequently, we consider an analytical method with compound Poisson distribution for sparse text data analysis. We divide the given data into zero and non-zero parts, and apply the proposed data analysis method suitable for each. Finally, a completed model is constructed by mixing the models applied to each of the two parts. To show the performance of our proposed method, we make experiments using simulated data and practical text data of patent documents. In our experimental results, we use Akaike information criterion (AIC) to evaluate the model performance between the proposed model and traditional models such as linear regression model, generalized linear model and zero-inflated Poisson model. The model with smaller values of AIC has better performance. Furthermore, we show that the AIC value of the proposed method is smaller than other models. Thus, we verify the validity of our research.

The organization of this paper is based on six sections. In the first section, we illustrate the research needs, motivations, proposal descriptions and contributions of the paper. We survey the related works of this paper such as text data analysis and the sparsity with zero-inflated problem in Section 2. In the following Section 3, the theoretical explanation of the proposed method and the performance evaluation measures of the model used in the experiments are presented. Next, we show the improved performance and validity of our proposed method by the experimental results using simulated data and real patent documents in Section 4. In Section 5, we deal with a comprehensive discussion, such as the research goal, background, limitations and scalability of contributions of our results. Finally, in Section 6, we provide the conclusions and future research tasks obtained through the research and experimental results.

2. Related Works

Big data, which is explained by immense size and heterogeneous data types, continues to rapidly increase its influence in very diverse fields. Additionally, big data has connected to various applications, such as IoT or AI. Hajjaji et al. (2021) introduced a systematic review of the connection between big data and IoT applications [18]. They answered to six research questions of the relations between big data and IoT according to different applications. Of course, security issues such as intrusion detection in the connection between big data and the IoT are important problems to be considered [19]. In this paper, we consider text data as a type of big data, because text data is one of the main data types that make up big data. Research related to text data was performed in various big data fields [5,6,20,21]. Among them, patent big data was analyzed by various machine learning algorithms [20,21] because patent documents also consist mainly of text data. A patent contains patent title, abstract, inventors, claims and descriptions of developed technology [22,23]. In addition, the technology classification codes, citations, applied dates and figures are included in a patent document [22,23]. Gamba (2017) used the patent data to show the effect of intellectual property rights for domestic innovation in the pharmaceutical field [24]. The

patent data has been used for technology management in various domains. Therefore, we consider the patent documents as text data in this paper. In text document analysis, the methods of extracting meaningful keywords and documents are needed [25]. This is because it is important to understand the relationship between keywords located at the top in text analysis. The research on topic models that can derive topics from given documents is also being actively conducted [26]. The text document clustering is one of the most important approaches in text data analysis. In recent studies, document clustering of vector model for dimmensionality reduction and document embedding for sentiment analysis are conducted [27–29]. As the use of Python data language has become more common, the development of Python libraries for text data analysis is actively progressing [30].

In general, there are some problems in the preprocessing and analysis of text data [5,6]. One of them is the sparsity problem with zero-inflated data [8,10,13,14,31]. For text data analysis using statistics and machine learning algorithms, we have to transform the text documents into a document-word matrix. This is because the data analysis methods provided by statistics and machine learning require structured data in the form of tables in which rows and columns consist of observations and variables, respectively. The document-word matrix is one of the structured data types. The document-word matrix is one of the structured datasets. The row and column of this matrix are document and word, respectively, and each element is the frequency value of a word occurring in a patent document. So far, various data analysis methods have been studied for the analysis of document-word matrix [5,32,33]. Kim and Jun built the patent document-word matrix using text mining, and analyzed it by graphical causal inference and copula regression. Thus, they illustrated the technological relations between the technology keywords of the Apple company and the subsequent data [32]. Park et al. (2017) extracted the text data from patent documents and analyzed the matrix data using fuzzy learning algorithms [20]. Kim et al. (2017) selected the keywords from patent data and used the penalized regression models for a patent keyword analysis [33]. Park and Jun (2020) analyzed patent keyword data using a technological cognitive diagnosis model [21]. They showed a hybrid model with statistics and cognitive science for text data analysis. Therefore, we need new and advanced models combined by interdisciplinary approaches for an efficient and effective text data analysis. As the size of the data increases, we encounter various difficulties in text data analysis. Data sparsity is one of these problems. The sparseness occurs in the preprocessing of text documents to make the structured data. That is, many elements of the document-word matrix have a value of zero. To solve this problem, Jun et al. (2014) provided a document-clustering method using dimension reduction and support vector clustering [7]. They tried to solve the sparsity problem in the given data by reducing the dimension of the given matrix data. Kim and Jun (2015) used the zero-inflated Poisson and negative binomial models for overcoming the sparseness in patent document data [32]. The proposed method tries to solve the sparsity problem by modeling the zero and non-zero parts of the given data, respectively. However, as the ratio of zeros included in the data increases, the existing methods for solving the sparsity problem show limitations. Therefore, we study one new method to solve the sparsity with zero-inflated problem in text data analysis.

There are various methods for zero-inflated data analysis [8–14]. Among them, the zero-inflated Poisson (ZIP) regression is a popular model for analyzing zero-inflated data [8,11,12]. The ZIP model is defined as (1) [9,11,34]:

$$\mathrm{P}(\mathrm{X}=\mathrm{x}) = \begin{cases} \pi + (1-\pi)e^{-m}, & \mathrm{x}=0 \\ (1-\pi)\frac{e^{-m}m^{x}}{x!}, & \mathrm{x}\geq 1 \end{cases} \tag{1}$$

ZIP separates the probability into two parts: zeros and non-zeros. In (1), the parameters are π and m. π is the proportion of zeros and m is the parameter of Poisson distribution. That is, π is the probability of $x = 0$ of binomial distribution and $(1 - \pi)$ is the probability

of $x \neq 0$ of Poisson distribution. ZIP is a regression model with two parameters: π and m. The mean of X is defined as (2) [9,34]:

$$E(X) = (1 - \pi)m \tag{2}$$

That is, the expectation of X is calculated by multiplying the parameter m by $(1 - \pi)$. Additionally, the variance of X is represented by (3) [9,34]:

$$Var(X) = (1 - \pi)\left(m + \pi m^2\right) \tag{3}$$

In (2) and (3), the probability π can be a constant or, depending on x and according to the binary outcome, zero may or may not occur. Using this, we can obtain the maximum likelihood estimations of m and π [9,34]. In addition, various models such as zero-inflated negative binomial (ZINB) are used for the analysis of zero excess data in diverse fields [11,12].

3. Proposed Method

3.1. Preprocessing of Text Data

In this paper, we propose a method for sparse text data analysis using compound Poisson linear modeling. The sparsity problem inevitably occurs in the preprocessed text data [7,8,10]. The preprocessing is performed by various text mining techniques [5,6]. We use R data language and its text mining packages for our text mining [6,35]. Table 1 shows the transformation functions used in our preprocessing of text documents [6].

Table 1. Transformation functions of tm package for text mining.

Transformation of tm Package	Function
tm_map(x,tolower)	Convert uppercases to lowercases
tm_map(x,removeNumbers)	Remove numbers
tm_map(x,removePunctuation)	Remove punctuations
tm_map(x,removeWords,stopwords("english"))	Remove stop-words
tm_map(x,stripWhitespace)	Remove unnecessary spaces
tm_map(x,removeWords, c(word list))	Remove user defined words (word list)

In Table 1, x is a corpus object that is a text document collection. For text data analysis based on statistics and machine learning algorithms, the text documents have to be transformed to a structured form, such as document-word matrix [5,6]. The matrix consists of documents and words for its rows and columns, respectively, and each element of the matrix represents the frequency of word occurred in a document. Figure 1 shows the matrix structure.

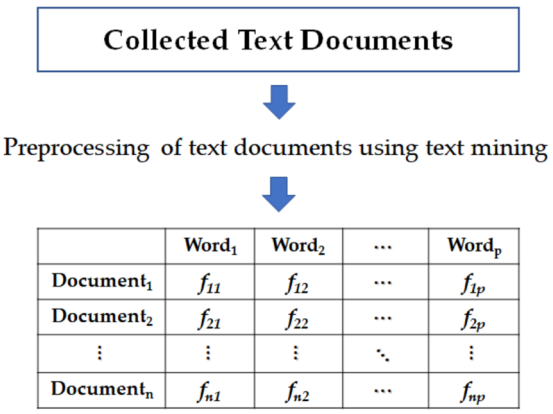

Figure 1. Document-word matrix.

In text data analysis, we first search the documents representing the target domain from various data sources. The document-word matrix is composed of n documents and p words, and the element f_{ij} is the frequency of jth word in ith document.

3.2. Proposed Model

Most elements of the document-word matrix have zero values because the number of words is much larger than the number of documents. That is, a word that appears even once in the whole document becomes one column in this matrix. In addition, the matrix has a frequency value only in the corresponding row and has a value of 0 in all other rows. Hence, the matrix data is a very sparse structure. This reduces the performance of predictive models. In previous studies, various models were proposed to solve the zero-inflated problem [9,11,12]. One of them is the zero-inflated count model such as zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB) models [9]. However, when the ratio of zeros included in the matrix becomes extremely high, the ZIP and ZINB models show the limitations to solve the sparsity problem. So, we propose a method to overcome the sparsity of text data using the compound Poisson linear model. In the proposed method, the matrix element is represented by (4):

$$Y = 0 \text{ or } Y \in [1, \infty) \tag{4}$$

where Y is an element value in the matrix. If the word does not occur in document, the value of Y is 0. Theoretically, the range of possible frequency values (matrix elements) is from 1 to infinity. Since the variable Y takes so many numbers from 0 to infinity, we consider Y as a continuous variable in our study.

In this paper, we use the keywords as the dependent and independent variables of statistical models. Additionally, we select a response variable from the keywords. The response variable is the target representing the domain of text documents. All variables other than the response can be predictors. We apply the compound distributions to analyze the sparse data with zero-inflated problem [14,17,36–41]. In the compound distribution, a random variable Y is defined as (5) [17]:

$$Y = \sum_{i=1}^{S} T_i \tag{5}$$

where S is frequency value of occurred keyword distributed to the discrete probability distribution in (6) [14,17,42]:

$$S \sim Poisson(m), \ (m > 0) \tag{6}$$

That is, S follows Poisson distribution with parameter m. In addition, T_i is the magnitude of ith keyword and distributed to gamma distribution as (7) [14,17,42]:

$$T_i \sim Gamma(\alpha, \beta), \ (\alpha > 0, \ \beta > 0) \tag{7}$$

In (7), α and β are shape and scale parameters of gamma distribution. Additionally, the random variable T_i has the property of independent and identically distributed (iid). In (6) and (7), S and T_i are independent of each other. Using (5), (6) and (7), we perform a sparse text data analysis using compound Poisson generalized linear mixed model (GLMM). If Y follows the compound Poisson distribution belonging to the exponential dispersion model (EDM), then it is distributed to the probability density function in (8) [14]:

$$\mathrm{f}(Y = y|\mu, \sigma) = d_p(y, \phi) \exp\left(\frac{y\theta - c(\theta)}{\phi}\right), \ \phi > 0 \tag{8}$$

where ϕ and θ are dispersion and canonical parameters, $d_p(y, \phi)$ is a probability density function of EDM. Additionally, $c(\theta)$ is a cumulant function similar to (9) [14]:

$$c(\theta) = \begin{cases} (((1-p)\theta + 1)^{\frac{2-p}{1-p}} - 1)/(2-p) & , p \neq 2 \\ -\log(1-\theta) & , p = 2 \end{cases} \quad (9)$$

In (8) and (9), the compound Poisson distribution has p greater than 1 and less than 2. The GLMM is an extended version of GLM and includes an additional term and random effect as (10) [17,39,41,43]:

$$Y = X\beta + Zb + e, \; E(b) = E(e) = 0 \quad (10)$$

where X and Z are design matrices. Additionally, β and b are fixed and random effect parameters, respectively. The difference between a fixed effect and a random effect is whether there is an assumption about the distribution of parameters or not. When making an estimate, the fixed effect explicitly estimates the parameter, and in the case of the random effect, the distribution of the parameter is obtained. The e is the error term, and the expectations of the random effect parameter and error term are all zeros. Based on the GLMM, we separate the frequency Y into two parts: zeros and non-zeros as (11) [17]:

$$Y_i = \begin{cases} 0 & , w_i \\ Compound\ Poisson\ distribution(\mu_i, \phi, p) & , 1 - w_i \end{cases} \quad (11)$$

where, w_i is the probability of weight when the frequency of the keyword is zero. Using (10) and (11), we analyze the sparse text data by compound Poisson GLMM. This is a generalized linear model mixture of degenerate and continuous distribution. In our research, we need a compound model to explain two parts of Y in (1). Therefore, we propose an approach for sparse text analysis using compound Poisson GLMM.

3.3. Model Evaluation and Procedure of Proposed Method

To evaluate the performance between comparative models, we use an AIC (Akaike information criterion) measure. AIC is a popular metric to evaluate performances between statistical models. The formula of AIC is defined as follows [43–45]:

$$\text{AIC} = -2\log\left(p\left(y|\hat{\theta}_{MLE}\right)\right) + 2k \quad (12)$$

In (12), k is the number of model parameters, $\hat{\theta}_{MLE}$ is the maximum likelihood estimator (MLE) of parameter θ, and y represents observed data. We have to select the model that minimizes AIC value. In Figure 2, we show the proposed method with four steps.

In Step 1, we preprocess the text documents using text-mining techniques and construct document-word matrix for text data analysis. The matrix has a very sparse data structure with excess zeros. Therefore, we separate the zero-inflated data into two parts of zero and non-zero in Step 2. We also use compound Poisson distribution to analyze the separated data. In the next step, we apply GLMM to compound Poisson distribution and build compound Poisson GLMM for sparse text data analysis. We overcome the sparsity problem of the preprocessed text data using compound Poisson GLMM in Step 4. Lastly, we carry out the sparse text data analysis and provide a better performance compared to generalized linear models. Thus, we compare AIC results of compound Poisson GLMM with linear regression in the next section.

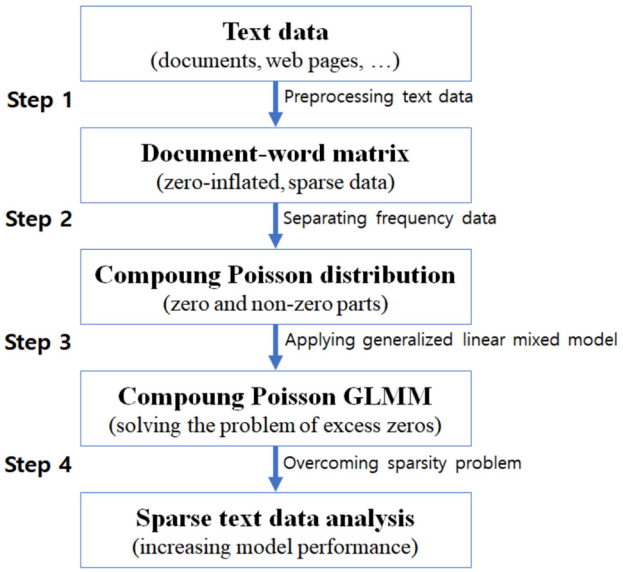

Figure 2. Procedure of proposed method.

4. Experimental Results

4.1. Simulation Data Analysis

First, we analyze a simulation data to illustrate the improved performance of our proposed method. We generated the random numbers from the generalized Poisson distribution with (0.7, 0.3, 0.4) and (0.352, 0.265, 0.342) for rate and dispersion parameters. The number of variables and the sample size are 3 and 10,000, respectively. So, we constructed the intermediate correlation matrix as follows.:

$$\begin{pmatrix} 1 & 0.4977 & 0.3671 \\ 0.4977 & 1 & 0.5049 \\ 0.3671 & 0.5049 & 1 \end{pmatrix}$$

We used the first variable as response variable y and others as explanatory variables x_1 and x_2. In this paper, we used the R data language and its packages for the simulation data generation [46,47]. Table 2 shows summary statistics of all variables.

Table 2. Summary statistics of simulation data.

Variable	Min	Q1	Median	Q3	Max	Mean	Percentage of Zero (%)
Y	0.0000	0.0000	1.0000	1.0000	7	0.7365	49.44
X_1	0.0000	0.0000	0.0000	1.0000	4	0.3028	74.08
X_2	0.0000	0.0000	0.0000	1.0000	5	0.4224	66.27

From the results in Table 2, we found that the simulation data is zero-inflated, for example, the zero percentage of x_1 is 74.08%. Using the variables, we constructed an analytical model as (13):

$$y = f(x_1, x_2) + \varepsilon \tag{13}$$

In the comparative models, we used y and (x_1, x_2) as response and explanatory variables. Table 3 illustrates the analysis results of comparative models using the simulation data of Table 2.

Table 3. Analysis results of comparative models using simulation data.

Explanatory Variable	LM	GLM	ZIP	CP
X_1	0.5195 (0.0001)	0.4975 (0.0001)	0.4453 (0.0001)	0.4981 (0.0001)
X_2	0.2104 (0.0001)	0.2269 (0.0001)	0.1712 (0.0001)	0.2274 (0.0001)
AIC	24,586.03	21,601.06	21,539.37	9087.50

In Table 3, we compared the compound Poisson model (CP) with comparative models which are linear regression model (LM), generalized linear model with Poisson distribution (GLM) and zero-inflated Poisson model (ZIP). We represented the estimated coefficients and p-values of x_1 and x_2 according to the comparative models. Additionally, we showed the AIC value of each model to evaluate model performance between compared models. We found that the AIC value of CP is the smallest among the comparison models. Therefore, we verified the better performance of CP for zero-inflated data analysis. Next, we used patent documents as real data to show the validity of our proposed method.

4.2. Patent Data Analysis

To show the performance of our proposed method, we used the patent document data related to drone technology. We searched the patent documents from the world patent database, the United States Patent and Trademark Office (USPTO) and the Korea Intellectual Property Rights Information Service (KIPRIS) [48,49]. The number of searched patents is 60,311 and we selected title and abstract from the patent documents. Using the preprocessing by text-mining techniques of Table 1, we built the document-word matrix as in Figure 3.

	unmanned	aerial	control	devic	system	machin	flight	bod	aircraft	connect	wing	power	rotor	arrang	data
1	0	0	0	0	1	0	0	0	1	4	1	0	0	0	0
2	0	0	0	0	1	0	0	0	1	4	1	0	0	0	0
3	0	0	0	2	3	0	1	0	6	0	3	1	0	2	0
4	0	8	0	3	0	0	0	0	0	3	0	0	0	1	0
5	0	8	0	3	0	0	0	0	0	3	0	0	0	1	0

	information	image	motor	landing	position	time	camera	air	signal	ground	batteri	bottom	shaft	communicate
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	2	0	0	0	0	0	0
4	0	0	0	0	0	0	2	0	3	0	0	1	0	0
5	0	0	0	0	0	0	2	0	3	0	0	1	0	0

	sensor	fuselag	transmission	propel	plane	wireless	speed	drone	electric	detect	rotat	direct	monitor	area
1	0	4	0	0	0	0	0	0	0	0	0	0	0	0
2	0	4	0	0	0	0	0	0	0	0	0	0	0	0
3	0	3	0	1	0	0	2	0	0	0	2	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0

	remot	station	flyin	drive	gear	water	engine	charg	storag	automat	energ	circuit	tail	efficienc	navi	video
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0

Figure 3. Part of the document-word matrix data of drone patent documents.

This figure shows a part of the entire preprocessed data. The row and column represent patent document and keyword, respectively. Each element of this matrix is the frequency value of keyword in a patent document. In Figure 3, we found that most elements are zero values. Hence, we knew that this matrix has a very sparse structure. In the next sub-section, we illustrate how the proposed method overcomes the sparsity and analyzes the matrix data. From the matrix, we selected the top 20 words with high frequencies for the keywords related to drone technology. The keywords were used as the variables in the proposed model. To check the sparsity of the document-keyword matrix, we show the summary statistics and percentage of zero values of the keywords in Table 4.

Table 4. Summary statistics of patent data.

Keyword	Min	Q1	Median	Q3	Max	Mean	Percentage of Zero Values
drone	0.0000	0.0000	0.0000	0.0000	28	0.1750	0.9518
control	0.0000	0.0000	0.0000	0.0000	33	1.4280	0.5806
device	0.0000	0.0000	0.0000	0.0000	24	0.9813	0.6546
flight	0.0000	0.0000	0.0000	0.0000	27	0.7134	0.7259
aircraft	0.0000	0.0000	0.0000	0.0000	26	0.6868	0.7748
wing	0.0000	0.0000	0.0000	0.0000	34	0.4832	0.8810
power	0.0000	0.0000	0.0000	0.0000	26	0.4548	0.8254
data	0.0000	0.0000	0.0000	0.0000	22	0.4104	0.8610
rotate	0.0000	0.0000	0.0000	0.0000	24	0.3800	0.8432
motor	0.0000	0.0000	0.0000	0.0000	19	0.3727	0.8650
drive	0.0000	0.0000	0.0000	0.0000	22	0.2996	0.8742
camera	0.0000	0.0000	0.0000	0.0000	27	0.2807	0.8928
signal	0.0000	0.0000	0.0000	0.0000	21	0.2724	0.8973
detect	0.0000	0.0000	0.0000	0.0000	25	0.2619	0.8913
battery	0.0000	0.0000	0.0000	0.0000	26	0.2320	0.9284
sensor	0.0000	0.0000	0.0000	0.0000	16	0.2266	0.9069
shaft	0.0000	0.0000	0.0000	0.0000	20	0.2160	0.9158
propel	0.0000	0.0000	0.0000	0.0000	24	0.2072	0.9229
automat	0.0000	0.0000	0.0000	0.0000	17	0.1669	0.9049
charge	0.0000	0.0000	0.0000	0.0000	28	0.1625	0.9606
remote	0.0000	0.0000	0.0000	0.0000	21	0.1512	0.9274

In Table 4, we represent minimum (Min), first quartile (Q1), median, third quartile (Q3), maximum (Max) and mean of frequency values. In addition, we show the percentage of zero values according to the keywords. In all the keywords, the values of Min, Q1, median and Q3 are all zeroes. In the case of the keyword *control*, although it has the largest average value, its mean value is only 1.4280. Additionally, the percentages of zero values of all keywords are very large. For example, the keyword *drive* contains 87.42% zeroes. Thus, we found that the data used in this experiment are very sparse. In general, the problem of sparsity reduces the performance of predictive models such as linear regression model [7,8]. Therefore, we performed zero-inflated patent data analysis using compound Poisson linear modeling. Table 5 illustrates the comparison results of performances between comparative models.

Table 5. Analysis results of comparative models using patent data.

Explanatory Variable	LM	GLM	ZIP	CP
control	−0.0047 (0.0106)	−0.0242 (0.0001)	0.0495 (0.0001)	−0.0259 (0.0509)
device	−0.0076 (0.0005)	−0.0503 (0.0001)	−0.0355 (0.0001)	−0.0565 (0.0007)
flight	−0.0023 (0.3590)	−0.0118 (0.0590)	0.0417 (0.0001)	−0.0181 (0.3271)
aircraft	−0.0262 (0.0001)	−0.4066 (0.0001)	−0.0776 (0.0001)	−0.2913 (0.0001)
wing	−0.0096 (0.0001)	−1.1221 (0.0001)	−0.0178 (0.1172)	−0.0991 (0.0001)
power	−0.0122 (0.0001)	−0.1155 (0.0001)	0.0107 (0.3166)	−0.0887 (0.0011)
data	0.0153 (0.0001)	−0.0528 (0.0001)	0.0011 (0.8683)	0.0592 (0.0007)
rotate	−0.0133 (0.0003)	−0.1484 (0.0001)	0.0160 (0.3388)	−0.1455 (0.0002)
motor	−0.0166 (0.0001)	−0.1615 (0.0001)	−0.0513 (0.0044)	−0.1289 (0.0002)
drive	−0.0128 (0.0013)	−0.1493 (0.0001)	−0.0515 (0.0017)	−0.0751 (0.0463)
camera	−0.0054 (0.1530)	−0.0326 (0.0007)	−0.0031 (0.7927)	−0.0277 (0.3116)
signal	0.0154 (0.0001)	0.0632 (0.0001)	0.0047 (0.5774)	0.0816 (0.0002)
detect	−0.0047 (0.2488)	−0.0313 (0.0015)	−0.0184 (0.1340)	−0.0168 (0.5493)
battery	−0.0139 (0.0001)	−0.1312 (0.0001)	−0.0220 (0.1689)	−0.1398 (0.0001)
sensor	0.0121 (0.0085)	0.0540 (0.0001)	0.0543 (0.0001)	0.0654 (0.0165)

Table 5. *Cont.*

Explanatory Variable	LM	GLM	ZIP	CP
shaft	−0.0160 (0.0008)	−0.3207 (0.0001)	0.0184 (0.5301)	−0.3351 (0.0001)
propel	−0.0013 (0.7699)	0.0109 (0.3811)	0.0053 (0.7493)	0.0085 (0.8046)
automat	−0.0164 (0.0109)	−0.1042 (0.0001)	0.0093 (0.6419)	−0.1148 (0.0289)
charge	0.0048 (0.2000)	0.0281 (0.0005)	0.0354 (0.0001)	0.0299 (0.2274)
remote	0.0071 (0.2329)	0.0325 (0.0120)	−0.0163 (0.3042)	0.0419 (0.2808)
AIC	172,438.10	74,791.42	36,524.01	35,501.73

In Table 5, the linear regression model is defined as (9):

$$drone = \beta_0 + \beta_1 control + \cdots + \beta_{20} remot + \varepsilon \quad (14)$$

where $(\beta_0, \beta_1, \ldots, \beta_{20})$ are regression coefficients and ε is the error term distributed to Gaussian with zero mean and equal variance. We compared the analysis results between CP and the other models according to model coefficient, its p-value and AIC. In Table 5, we knew that the AIC value of CP is smaller than the comparative models, LM, GLM and ZIP. This means that the model explanatory power of CP is greater than the other models. From the results in Tables 4 and 5, we could confirm the improved performance of the CP model in sparse text data analysis.

5. Discussion

In this paper, we used the compound Poisson distribution to build the model for zero-inflated patent data analysis. We separated the zero-inflated data to the parts of zero and non-zero. This is similar to the traditional zero-inflated models such as ZIP or ZINB. However, the CP is different from the existing zero-inflated models in two respects. First, the proposed method applied the degenerated distribution to the zero values. Second, the continuous distribution was used for non-zero values in our method. Since each element value of our patent document-keyword matrix is an integer ranging from 0 to infinity, we consider apply a continuous probability distribution to this frequency. This is because the range of keyword frequency values is infinite.

The goal of this paper is to construct a model for solving the zero-inflated problem in patent data analysis. We also tried to solve the problem with the different approach from existing zero-inflated data analysis methods such as ZIP and ZINB. That is, the probability distributions are applied to both the zero and non-zero parts. Our research contributes to the text data analysis in diverse fields, such as text documents, web contents or social network service (SNS) posts, as well as patent document data. For example, our research can be extended and applied to the analysis of data generated by the IoT. Numerous sensing data generated in the IoT environment also have the zero-inflated problem. This is because most of the data observed through the sensor system are the values without change; that is, zero values. Therefore, we expect that the CP model proposed in this paper will provide an efficient and improved performance compared to other existing models with regard to the zero-inflated problem within obtained data, while many other researchers conduct various studies in their respective fields.

6. Conclusions

A large part of big data consists of text documents. Thus, many studies related to text data analysis have been conducted and introduced. Patent document is one of the most popular text data. In this paper, we focused on the patent data analysis. For patent data analysis based on statistics and machine-learning algorithms, we preferentially transform patent documents into the structured data using text-mining techniques. In general, the preprocessed data is a matrix form consisting of patent document and keyword for row and column, respectively. Most of the elements of this matrix are zeros, so we run into

the sparsity with zero-inflated problem in the document-word matrix. This is an obstacle to analyze patent text data. We have to solve this problem for an efficient patent data analysis. To overcome the zero-inflated problem, we proposed a method of patent analysis using compound Poisson modeling. In this paper, we separated the sparse data to zero and non-zero parts. Additionally, we applied the proposed method to each of the two parts. For our experiments to evaluate the model performance between our proposed method and traditional models, we performed experiments using simulated data and real patent documents. First, we found that the AIC value of CP is the smallest in the compared models, so, we verified the improved performance of our method. Second, we searched the patent documents related to drone technology from the world patent databases and preprocessed them to construct a patent document-keyword matrix. This matrix has an extreme sparsity problem with most of its elements having zero values. Similar to the simulation data, we compared the AIC value of CP with LM, GLM and ZIP, and the value of CP was smaller than others. Therefore, we confirm that the model performance of CP is better than other comparative models. As the use of AI increases, the importance of big data analysis for learning from data increases. The analysis of text data is absolutely necessary in natural language processing, which is a core technology of AI. Therefore, we will carry out the research on methods to solve various problems arising from text data analysis as well as the zero-inflated problem. In our future work, we will study on more advanced models based on statistics and machine-learning algorithms to solve the sparsity problems in text data analysis. In addition, we will try to combine Bayesian inference and generative deep learning to generate synthetic data for zero-inflated data augmentation. The synthetic data is replaced with the original zero-inflated data and is used for text data analysis. Furthermore, we will contain various visualization methods, such as plots, graphs or graphical models, in our next research.

Author Contributions: S.P. designed this research and collected the dataset for the experiment. S.J. analyzed the data to show the validity of this paper and wrote the paper and performed all the research steps. In addition, all authors have cooperated in revising the paper. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2020R1I1A3A04037885).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

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

References

1. Arijanto, J.E.; Geraldy, S.; Tania, C.; Suhartono, D. Personality Prediction Based on Text Analytics Using Bidirectional Encoder Representations from Transformers from English Twitter Dataset. *Int. J. Fuzzy Log. Intell. Syst.* **2021**, *21*, 310–316. [CrossRef]
2. Kim, S.; Son, D.; Park, M.; Hwang, H. Developing a Big Data Analytic Model and a Platform for Particulate Matter Prediction: A Case Study. *Int. J. Fuzzy Log. Intell. Syst.* **2019**, *19*, 242–249. [CrossRef]
3. Lee, J.; Lee, J. Constructing Efficient Regional Hazardous Weather Prediction Models through Big Data Analysis. *Int. J. Fuzzy Log. Intell. Syst.* **2016**, *16*, 1–12. [CrossRef]
4. Zolkepli, M.; Dong, F.; Hirota, K. Automatic Switching of Clustering Methods based on Fuzzy Inference in Bibliographic Big Data Retrieval System. *Int. J. Fuzzy Log. Intell. Syst.* **2014**, *14*, 256–267. [CrossRef]
5. Feinerer, I.; Hornik, K.; Meyer, D. Text mining infrastructure in R. *J. Stat. Softw.* **2008**, *25*, 1–54. [CrossRef]
6. 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.
7. 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]
8. Kim, J.; Jun, S. Zero-Inflated Poisson and Negative Binomial Regressions for Technology Analysis. *Int. J. Softw. Eng. Its Appl.* **2016**, *10*, 431–448. [CrossRef]
9. Cameron, A.C.; Trivedi, P.K. *Regression Analysis of Count Data*, 2nd ed.; Cambridge University Press: Cambridge, UK, 2013.

10. Feng, C.X. A comparison of zero-inflated and hurdle models for modeling zero-inflated count data. *J. Stat. Distrib. Appl.* **2021**, *8*, 8. [CrossRef]
11. Hilbe, J.M. *Negative Binomial Regression*, 2nd ed.; Cambridge University Press: Cambridge, UK, 2011.
12. Hilbe, J.M. *Modeling Count Data*; Cambridge University Press: Cambridge, UK, 2014.
13. Dencks, S.; Piepenbrock, M.; Schmitz, G. Assessing Vessel Reconstruction in Ultrasound Localization Microscopy by Maximum Likelihood Estimation of a Zero-Inflated Poisson Model. *Proc. IEEE Trans. Ultrason. Ferroelectr. Freq. Control* **2020**, *67*, 1603–1612. [CrossRef]
14. Hwang, H.; Song, E.; Park, N.; Lee, W. Analyzing Precipitation Data with Zeroes Using Compound Poisson Distribution. *J. Korean Data Anal. Soc.* **2016**, *18*, 129–140.
15. Sert, O.C.; Sahin, S.D.; Özyer, T.; Alhajj, R. Analysis and prediction in sparse and high dimensional text data: The case of Dow Jones stock market. *Phys. A: Stat. Mech. Its Appl.* **2020**, *545*, 123752. [CrossRef]
16. Unnikrishnan, P.; Govindan, V.K.; Madhu Kumar, S.D. Enhanced sparse representation classifier for text classification. *Expert Syst. Appl.* **2019**, *129*, 260–272.
17. Zhang, Y. *Package 'cplm' ver. 0.7-10, Likelihood-Based and Bayesian Methods for Various Compound Poisson Linear Models*; CRAN of R Project; R Foundation for Statistical Computing: Vienna, Austria, 2022.
18. Hajjaji, Y.; Boulila, W.; Farah, I.R.; Romdhani, I.; Hussain, A. Big data and IoT-based applications in smart environments: A systematic review. *Comput. Sci. Rev.* **2021**, *39*, 100318. [CrossRef]
19. Javanmardi, S.; Shojafar, M.; Mohammadi, R.; Persico, V.; Pescapè, A. S-FoS: A secure workflow scheduling approach for performance optimization in SDN-based IoT-Fog networks. *J. Inf. Secur. Appl.* **2023**, *72*, 103404. [CrossRef]
20. Park, S.; Lee, S.; Jun, S. Patent Big Data Analysis using Fuzzy Learning. *Int. J. Fuzzy Syst.* **2017**, *19*, 1158–1167. [CrossRef]
21. Park, S.; Jun, S. Technological Cognitive Diagnosis Model for Patent Keyword Analysis. *ICT Express* **2020**, *6*, 57–61. [CrossRef]
22. Hunt, D.; Nguyen, L.; Rodgers, M. *Patent Searching Tools & Techniques*; Wiley: Hoboken, NJ, USA, 2007.
23. 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.
24. Gamba, S. The effect of intellectual property rights on domestic innovation in the pharmaceutical sector. *World Dev.* **2017**, *99*, 15–27. [CrossRef]
25. Truica, C.; Darmont, J.; Boicea, A.; Radulescu, F. Benchmarking top-k keyword and top-k document processing with T2K2 and T2K2D2. *Future Gener. Comput. Syst.* **2018**, *85*, 60–75. [CrossRef]
26. Truica, C.; Radulescu, F.; Boicea, A. Comparing Different Term Weighting Schemas for Topic Modeling. In Proceedings of the 2016 18th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC), Timisoara, Romania, 24–27 September 2016; pp. 307–310.
27. Radu, R.; Radulescu, I.; Truica, C.; Apostol, E.; Mocanu, M. Clustering Documents using the Document to Vector Model for Dimensionality Reduction. In Proceedings of the 2020 IEEE International Conference on Automation, Quality and Testing, Robotics (AQTR), Cluj-Napoca, Romania, 21–23 May 2020; pp. 1–6.
28. Radulescu, I.; Truica, C.; Apostol, E.; Boicea, A.; Mocanu, M.; Popeanga, D.; Radulescu, F. Density-based Text Clustering using Document Embeddings. In Proceedings of the 36th IBIMA Conference, Granada, Spain, 4–5 November 2020; pp. 6222–6230.
29. Mitroi, M.; Truica, C.; Apostol, E.; Florea, A. Sentiment Analysis using Topic-Document Embeddings. In Proceedings of the IEEE 16th International Conference on Intelligent Computer Communication and Processing, Cluj-Napoca, Romania, 3–5 September 2020; pp. 75–82.
30. Truica, O.; Aostol, E.; Paschke, A. Awakened at CheckThat! 2022: Fake news detection using BiLSTM and sentence transformer. In Proceedings of the Conference and Labs of the Evaluation Forum, Bologna, Italy, 5–8 September 2022; pp. 1–9.
31. Altay, A.; Baykal-Gürsoy, M. Imperfect rail-track inspection scheduling with zero-inflated miss rates. *Transp. Res. Part C* **2022**, *138*, 103608. [CrossRef]
32. 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]
33. Kim, J.; Ryu, J.; Lee, S.; Jun, S. Penalized Regression Models for Patent Keyword Analysis. *Model Assist. Stat. Appl.-Int. J.* **2017**, *12*, 239–244. [CrossRef]
34. Wagh, Y.S.; Kamalja, K.K. Zero-inflated models and estimation in zero-inflated Poisson distribution. *Commun. Stat. -Simul. Comput.* **2018**, *47*, 2248–2265. [CrossRef]
35. R Development Core Team. R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing. Available online: http://www.R-project.org (accessed on 1 March 2022).
36. Babai, M.Z.; Chen, H.; Syntetos, A.A.; Lengu, D. A compound-Poisson Bayesian approach for spare parts inventory forecasting. *Int. J. Prod. Econ.* **2021**, *232*, 107954. [CrossRef]
37. Haakonsson, S.; Rodríguez, M.A.; Carballo, C.; Pérez, M.D.C.; Arocena, R.; Bonilla, S. Predicting cyanobacterial biovolume from water temperature and conductivity using a Bayesian compound Poisson-Gamma model. *Water Res.* **2020**, *176*, 115710. [CrossRef]
38. Prak, D.; Teunter, R.; Babai, M.Z.; Boylan, J.E.; Syntetos, A. Robust compound Poisson parameter estimation for inventory control. *Omega* **2021**, *104*, 102481. [CrossRef]
39. Xie, J.; Zhang, Z. Statistical estimation for some dividend problems under the compound Poisson risk model. *Insur. Math. Econ.* **2020**, *95*, 101–115. [CrossRef]

40. Su, W.; Yong, Y.; Zhang, Z. Estimating the Gerber–Shiu function in the perturbed compound Poisson model by Laguerre series expansion. *J. Math. Anal. Appl.* **2019**, *469*, 705–729. [CrossRef]
41. Zhang, Y. Likelihood-based and Bayesian methods for Tweedie compound Poisson linear mixed models. *Stat. Comput.* **2013**, *23*, 743–757. [CrossRef]
42. Hogg, R.V.; Mckean, J.W.; Craig, A.T. *Introduction to Mathematical Statistics*, 8th ed.; Pearson: Essex, UK, 2020.
43. Murphy, K.P. *Machine Learning: A Probabilistic Perspective*; MIT Press: Cambridge, MA, USA, 2012.
44. Bruce, P.; Bruce, A.; Gedeck, P. *Practical Statistics for Data Scientists*, 2nd ed.; O'Reilly Media: Sebastopol, CA, USA, 2020.
45. 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.
46. Li, H.; Chen, R.; Nguyen, H.; Chung, Y.; Gao, R.; Demirtas, H. *Package 'RNGforGPD' Version 1.1.0, Random Number Generation for Generalized Poisson Distribution*; CRAN of R Project; R Foundation for Statistical Computing: Vienna, Austria, 2022.
47. Li, H.; Demirtas, H.; Chen, R. RNGforGPD: An R Package for Generation of Univariate and Multivariate Generalized Poisson Data. *R J.* **2020**, *12*, 173–188. [CrossRef]
48. USPTO. The United States Patent and Trademark Office. Available online: http://www.uspto.gov (accessed on 1 May 2022).
49. KIPRIS. Korea Intellectual Property Rights Information Service. Available online: www.kipris.or.kr (accessed on 1 March 2022).