International Journal of Computing Sciences Research (ISSN print: 2546-0552; ISSN online: 2546-115X)
Vol. 7, pp. 1162-1186
doi: 10.25147/ijcsr.2017.001.1.95
https://stepacademic.net

**Long Paper**

# Province of Laguna Legislative Management and Tracking System with the Application of Latent Dirichlet Allocation (LDA) Algorithm

Charles Alfred A. Cruz
College of Computer Studies
Laguna State Polytechnic University-Santa Cruz, Philippines
charlesalfredcruz@lspu.edu.ph
(corresponding author)

Francis F. Balahadia
College of Computer Studies
Laguna State Polytechnic University-Siniloan, Philippines
francis.balahadia@lspu.edu.ph

Recommended citation:

> Cruz, C. A., & Balahadia, F. F. (2023). Province of Laguna Legislative Management and Tracking System with the Application of Latent Dirichlet Allocation (LDA) Algorithm. *International Journal of Computing Sciences Research*, *7*, 1162-1186. https://doi.org/10.25147/ijcsr.2017.001.1.95

## Abstract

*Purpose* – The legislative branch of a province is generally in charge of making laws. In addition to this, they are also in charge of enacting programs and policies for the general well-being of the citizens within the province. The general public may not be able to keep track of legislative performances since there is a growing number of legislative-generated documents. This study developed a system that integrated the application of topic modeling in crafting ordinances, resolutions, and policies for the Province of Laguna.

*Method* – This research employed the SCRUM methodology process in the development of the system and Latent Dirichlet Allocation (LDA) topic modeling using R language to classify text within a document to a particular topic and created model through Observation-based evaluation and Quantitative metrics such as Perplexity and Coherence to determine k-value (number of topics) based on the corpus wherein it was collected in the Twitter and Legislative Management System Portal.

*Results* – The results showed the LDA used returned the optimal value for perplexity and coherence which was determined by testing different k-values ranging from 1 to 2 which is presented to users as a line graph. The developed system provided a system module

that can enable users to find the optimal number of topics (k value) and present the results in a visually appealing interface on the user's account portal which gives insights into what possible new ideas in formulation ordinances, resolutions, and policies in Laguna.

Conclusion – The developed system in this study allows legislators of the province of Laguna to collect public posts from the social networking site Twitter and use Latent Dirichlet Allocation (LDA) topic modeling. It also provides an interactive graph that allows users to explore the LDA model generated by the system and helps to reveal topics of concern from the community that leads to government officials in formulating policies and ordinances appropriate for the needs of the community.

*Recommendations* – It is recommended to develop an additional module that automatically generates the topic model based on the selected LDA evaluation procedure and should be tested in a larger-sized corpus to further test its capabilities as well as to improve the list of the stop words and noise removal feature.

*Implications* – The system can be used to simply accomplish the document trail page where users can preview document details and the application of the visualization techniques in the system helps to facilitate the to provide an impression by extracting words, and topics that can be a basis of crafting programs and priorities of the government officials in taking actions to the citizen concerns.

*Keywords* – topic modeling, LDA, tweets, legislations, public response, Laguna, Sangguniaang Panlalagiwan, word clouds

## INTRODUCTION

Following the Philippine Republic Act No. 7160 more commonly known as the "Local Government Code of 1991", The Sangguniang Panlalawigan (SP) also known as the Provincial Board shall exercise the local legislative power of a province with the vice governor service as the presiding officer. Laguna is one of the provinces in the Philippines whose powers are exercised through their SP. Along with the Provincial Vice-Governor serving as the presiding officer, three (3) Board Members from the 1st and 2nd District and two (2) Board Members from the 3rd and 4th District are included as members.

The legislative body is generally in charge of making laws. This includes the enactment of ordinances, approval of resolutions, and the appropriation of funds for the general well-being of the province and its citizens concerning the proper exercise of the corporate powers of the province. Aside from the review and approval of ordinances, resolutions, and funds, the Sangguniang Panlalawigan is also tasked with accrediting (R.A.

7160). The Sangguniang Panlalawigan of Laguna reviews and approved an average of 291 resolutions and ordinances yearly. In the year 2020, According to the Sangguniang Panlalawigan of Laguna, they reviewed and approved 146 ordinances submitted by cities and municipalities. Additionally, they have also approved and reviewed 148 Ordinances originating from the Sangguniang Panlalawigan board members.

The general public may not be able to keep track and monitor the legislative performance of each legislator due to the growing number of legislative-generated documents. (Lin, Chou, Liao and Hao, 2011). Locating and managing these files among several others can be considered a time-consuming and tedious process for administrative staff as well as the general public. Public access to these documents is an equally important matter. It is also important for citizens to track and be aware of Ordinances and Resolutions implemented in their hometowns. In addition to this, the linkage between government and citizens is strengthened when the public has the means to voice out their concerns and be heard by the legislature. In light of this, on a global scale, governments are providing online services to citizens over the internet through web portals (Zaidi & Qteishat, 2012).

Additionally, it is an important principle of a government system in a country such as the Philippines that public policies are decided upon by the citizens. Constitutional and Legal reforms have resulted in a national 'citizen charter' in which the citizens and citizen groups have been enabled to complain to government officials whenever there is inadequacy on the delivery of public services (Porio, 2017). Citizen participation is represented as a major component, as it is the way to bridge government decisions and the real expectations of citizens (Buccafurri, Fotia & Lax, 2015). This further supports the idea that the citizens are sources of power and that their opinions should at least be given consideration in the molding of actions the government is going to make. The Sangguniang Panlalawigan is no exception to these kinds of problems.

Some pieces of literature conducted citizen participation and engagement to governments. The study of Alguliyev et.al. (2019) discussed extracting the hidden social network using the analysis of the user's sentiments through opinion and text mining in improving the management of e-government services. Furthermore, Jelonek et.al (2020) the application of sentiment analysis can be a tool for building city development strategies and other possible projects for the city. Moreover, Hubert, et al. (2018) use several visualization techniques in assessing the communication of the community and the government in public concern issues through social media network like Twitter that reveals patterns and trend in government-citizen interactions. But this literature did not integrate document tracking, topic modeling and only focuses on getting the sentiments of the citizen.

Document tracking which pertains to the recording and monitoring the movement of documents has been a time-consuming task. Dislocation and overlooking of the timeline have always been problems in document control. An effective tool such as a web-based

system is the easiest way to be implemented in a workplace (Salleh, Ujir, Sapawi & Hashim, 2020). Because of the nature of web-based systems, they can be accessed online thus supporting employees' needs for a work-from-home scenario.

Documents processed and archived in these systems also have the potential to reveal abstract topics that could be useful information to legislators on which subject matter was heavily focused on and which have minimal attention from legislators. On this particular note, the use of social media by the Filipino people in terms of democratic consolidation is considered exceptionally remarkable because, during the last four years, The Philippines has been the social media leader in terms of the number of users worldwide. In terms of internet usage, they rank first globally with an average daily screen time of 10 hours with almost 50% of the adult population are using the internet (Yusingco, 2020; Pablo, 2018). This opens up an opportunity for the Sangguniang Panlalawigan to hone in and listen to the voice of their people to further improve and support their decision-making process in formulating and drafting legislation.

Topic modeling is an unsupervised machine learning technique that's capable of scanning a set of documents, detecting word and phrase patterns within them, and automatically clustering word groups and similar expressions that best characterize a set of documents (Pascual, 2019). Topic modeling could potentially be a tool for legislators to discover abstract topics from archived and processed legislative documents to identify heavily and lightly focused topics and also identify abstract topics from a collection of social media posts by the citizen to identify which topics of concern needs more attention for future actions.

This research aimed to develop a system that allows users to manage and process legislative documents submitted to the Sangguniang Panlalawigan of Laguna. Additionally, it also aimed at developing a system web module that allows legislators of the province of Laguna to collect public posts from the social networking site Twitter. It also aimed to provide users to apply Latent Dirichlet Allocation (LDA) topic modeling to reveal abstract topics from archived documents and social media posts collected using the Twitter scraper module. Since the number of k-value (topics to generate by LDA) is decided by the user, a system module that enables users to evaluate the topic model was also developed using different techniques such as Observation-Based evaluation and Quantitative metrics evaluation.

## METHODOLOGY

The goal of this study is to develop a system capable of handling and managing legislative documents as well as collecting public responses from the social networking site, Twitter. Additionally, it also aims to develop a system module that applies Latent Dirichlet Allocation (LDA) and different evaluation tools for LDA such as Observation-based evaluation and Quantitative metrics such as Perplexity and Coherence. The system

also aims to visualize topic modeling results using an R package called LDAvis to help users analyze and interpret topics from the topic model result.

## *DATA COLLECTION AND PREPROCESSING*

The developed system uses several data resources in and out of the system's environment. These data are classified by the researcher into two namely Collected User Contents and Sangguniang Panlalawigan Legislative Documents. Collected user contents are identified by public user posts in the social networking site Twitter and posts submitted to the forum module of the developed system. Sangguniang Panlalawigan Legislative Documents, on the other hand, are documents archived to the system by the system administrators. Collection of these data can provide an alternative way for Sangguniang researchers on the opinions, sentiments, and other topics hidden within the said data resources. These data are considered raw and unstructured which is why data Preprocessing is applied.

Preprocessing is one of the key components in a typical text classification framework. It is simply the process of transforming your text into a form that is ready for prediction and analysis for the task. Text preprocessing is traditionally an important step for natural language processing (NLP) tasks. It transforms text into a more digestible form so that machine learning algorithms can perform better (Jose B.K., 2021). However, data from social media and forum posts are different from formal legislative documents. For example, social media and forum posts might contain special characters that could also be analyzed and interpreted such as emojis and emoticons. Emojis are Unicode picture symbols, used as a shorthand to specify principles and ideas. Emojis on smartphones, in chat, and email packages have come to be extraordinarily famous worldwide.

For example, Instagram, an online mobile photo-sharing, video-sharing, and social networking platform, reported in March 2015 that nearly half of the texts on Instagram contained emojis (Dimson, 2015). The researcher developed a separate preprocessing module for collected user contents to preserve special characters such as emojis and emoticons before saving them in the data set. Figure 1 illustrates the model of data collection and preprocessing procedures implemented by the system to generate a dataset for the NLP modules.

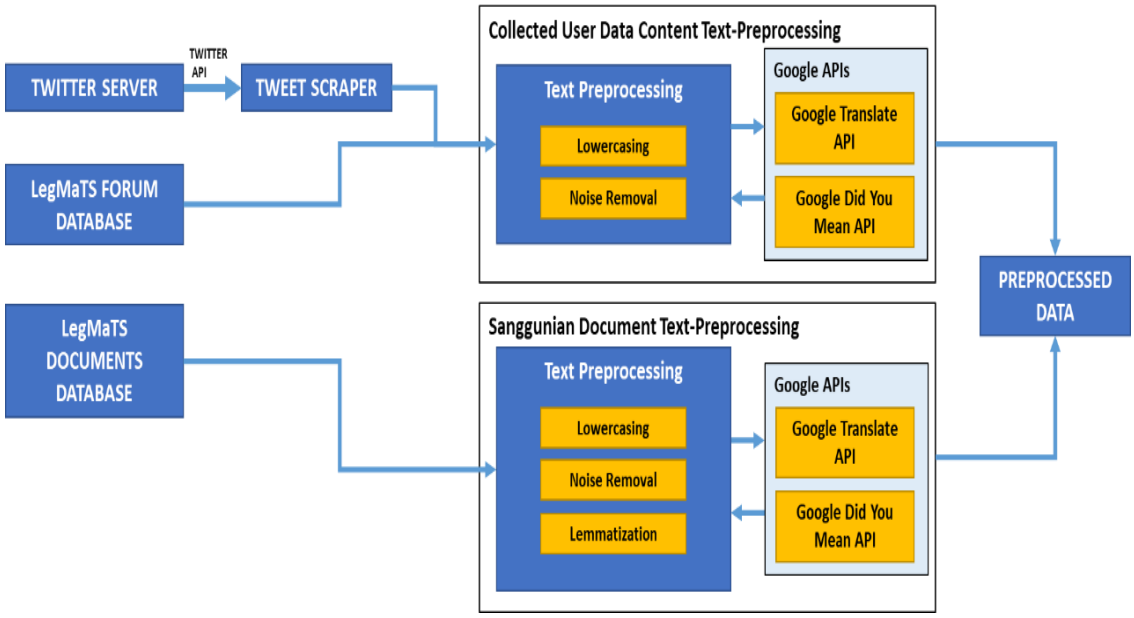


*Figure 1.* The Data Collection and Preprocessing stage.

**Twitter API and Scraping**

Social media can be a gold mine of data in regards to consumer sentiment. Social media platforms such as Twitter are readily available and hold useful information since their users may post opinions that are sometimes unfiltered which can be retrieved with ease (Beck, 2020).

To perform the collecting of tweets, a Twitter developer account was created and registered to interface with the Twitter Search API. Search results are then collected and saved through the tweet scraper module for preprocessing.

**Lowercasing**

Although commonly overlooked, the lowercasing of the text data is one of the most effective as well as simplest forms of text preprocessing. It is commonly applied to a majority of text mining as well as NLP problems. It can also help in The developed system implemented the strtolower() function to transform the data set into its lowercase form.

**Noise Removal**

Noise removal pertains to the process of removing a string's digits and other pieces of characters that could interfere with the text analysis process. Removal of special characters was carefully done especially to tweets and forum posts to not lose emojis and emoticons that also had sentiment values. URLs and tags were also detected by the developed system and were removed from the cleaning outcome.

**Lemmatization**

The purpose of Lemmatization is rather just like Stemming withinside the experience that their purpose is to get the root form of a phrase with the distinction being that lemmatization attempts to do it the right way. It does now no longer simply chop parts off, it transforms phrases to the actual root. For example, the phrase "better" might map to "good" (Senthilkumar, RubanRaja, & Monisha, 2021). The data set was then interfaced to a lemmatization library to decide the root of actual words from the data set.

**Stop Words Removal**

The developed system implemented stop words removal by removing commonly used words that hold no bearing in the generation of topics for LDA. Examples of stop words in English are "a", "the", "is", "are" etc. The researchers implemented the native stop words list provided by the topicmodels() function in R.

**N-gram**

N-grams are used to predict the existence of words through their N – 1 previous word. It is also used in text mining and NPL tasks in which a set of co-occurring words within a given window and when computing the n-grams you typically move one word forward (Ganesan, n.d). In this system, it used unigram because the system only one word can separate in the collected words in the social media to assess the common words raised by the citizen.

## *TOPIC MODELING APPROACH*

Topic modeling is a method used to perform unsupervised classification of documents, which holds similarity to the clustering on numeric data, which discovers some natural groups of items (in this case, topics) even when the user is not sure what they are looking for. The most frequently used topic modeling algorithm is the LDA or the Latent Dirichlet Allocation.

LDA is a topic modeling technique that describes the probability procedure of a document (Subeno & Kusumaningrum, 2018). LDA is a way of automatically discovering topics that a collection of sentences contains.
According to Wowchemy (2021) In LDA, it represents documents as a combination of topics that gives out words that contain certain probabilities. LDA assumes that these said documents are produced in the following fashion: When writing the documents, you.

- Decide on the number of words N the document will have. By default, the developed system returns 25 words related to each topic.
- Choose a topic mixture for the document (according to a Dirichlet probability distribution over a fixed set of K topics). The developed system asks the user to input the number of topics (K) and generates a user-friendly graph to easily visualize the topic model.
- Generate each word in the document by:
    - First picking a topic (according to the distribution that you sampled above;
    - Then using the topic to generate the word itself (according to the topic's multinomial distribution).

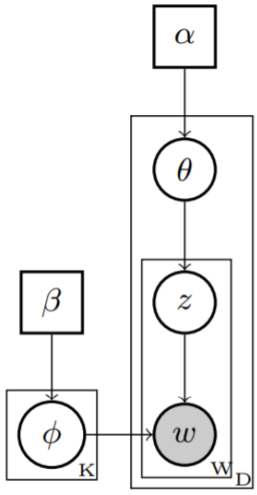


*Figure* 2. Graphic model for Latent Dirichlet Allocation.

In the study of Porteous, et al. (2008) we adopted the LDA model in the developed application. Wherein the LDA models each of D documents as a mixture over K latent topics, each of which describes a multinomial distribution over a W word vocabulary. Figure 2 shows the graphical model representation of the LDA model. The LDA model is equivalent to the following generative process for words and documents: For each of Nj words in document j

1. sample a topic zij ~ Multinomial(θj )
2. sample a word xij ~ Multinomial(φzij )

where the parameters of the multinomials for topics in a document θj and words in a topic φk have Dirichlet priors. Intuitively we can interpret the multinomial parameter φk as indicating which words are important in topic k and the parameter θj as indicating which topics appear in document j Given the observed words x = {xij}, the task of Bayesian inference is to compute the posterior distribution over the latent topic indices z = {zij}, the mixing proportions θj, and the topics φk. An efficient inference procedure is to use collapsed Gibbs sampling, where θ and φ are marginalized out, and only the latent

variables z are sampled. After the sampler has burned in we can calculate an estimate of θ and φ were given z.

The defined summations of the data by Nwkj = #{i : xij = w, zij = k}, and use the convention that missing indices are summed out, so that Nkj = P w Nwkj and Nwk = P j Nwkj . In words, Nwk is the number of times the word w is assigned to the topic k and Nkj is the number of times a word in document j has been assigned to topic k. Given the current state of all but one variable zij, the conditional probability of zij is then

$$p(z_{ij} = k|\mathbf{z}^{\neg ij}, \mathbf{x}, \alpha, \beta) = \frac{1}{Z} a_{kj} b_{wk}$$

*Equation 1*

where

$$a_{kj} = N_{kj}^{\neg ij} + \alpha \qquad b_{wk} = \frac{N_{wk}^{\neg ij} + \beta}{N_k^{\neg ij} + W\beta},$$

*Equation 2*

Z is the normalization constant

$$Z = \sum_k a_{kj} b_{wk},$$

*Equation 3*

and the superscript ¬ij indicates that the corresponding datum has been excluded in the count summations $N_{wkj}$.

**for** $i \leftarrow 1$ **to** $N$
**do**
$u \leftarrow$ draw from Uniform$[0, 1]$
**for** $k \leftarrow 1$ **to** $K$
**do**
$P[k] \leftarrow P[k-1] + \frac{(N_{kj}^{\neg ij}+\alpha)(N_{x_{ij}k}^{\neg ij}+\beta)}{(N_k^{\neg ij}+W\beta)}$
**for** $k \leftarrow 1$ **to** $K$
**do**
**if** $u < P[k]/P[K]$
**then** $z_{ij} = k, stop$

*Figure 3*. LDA Gibbs Sampling Algorithm.

Figure 3 shows the Gibbs Sampling Algorithm used by the developed system. An iteration of Gibbs sampling proceeds by drawing a sample for zij according to (1) for each word i in each document j. A sample is typically accomplished by first calculating the normalization constant Z, then sampling zij according to its normalized probability. Given

the value sampled for zij the counts Nkj, Nk, Nwk are updated. The time complexity for each iteration of Gibbs sampling is then O(NK) (Porteous et al.,2008).

Given a sample we can then get an estimate for ˆθj and φˆk were given z:

Results from the model were then visualized using the graphs and were loaded to LDAvis - a web-based interactive visualization of topics estimated using Latent Dirichlet Allocation that is built using a combination of R and D3.

### Word-topic probabilities

To get the word-topic probability, the researcher implemented a package in R called tidytext particularly the tidy() method. The tidytext package provides this method for extracting the per-topic-per-word probabilities, called β ("beta"), from the model.

```
beta_wide <- ap_topics %>%
  mutate(topic = paste0("topic", topic)) %>%
  pivot_wider(names_from = topic, values_from = beta) %>%
  filter(topic1 > .001 | topic2 > .001) %>%
  mutate(log_ratio = log2(topic2 / topic1))

beta_wide
write.csv(beta_wide,file=paste("C:/xampp/htdocs/LegMaTS/assets/transaction/
```

*Figure 4*. R Code for extracting the word-topic probability

The source code indicated in Figure 4 shows how to generate the word-topic probability applied in the system. A one-topic-per-term-per-row format output is derived from the model based on the code illustrated in Figure 4.

## *TOPIC MODELING EVALUATION*

### Observation-based Approach

The simplest way of evaluating a topic is to look at the most frequent words in the topic. This can be done in tabular form, for example, by listing the top 10 words in each topic, or other formats (Rabindranath, 2020). Aside from this, another visually appealing way to present these results are Word Clouds.

### Quantitative Metrics-based Approach

Although evaluation methods based on human judgment can outcome in good results, they are usually not cost-effective and also time-consuming to do. Additionally, human judgment isn't clearly defined and humans have a tendency to disagree on what makes a good topic.

**Perplexity**

The most common way to evaluate a probabilistic model is to measure the log-likelihood of a held-out test set (Jeong, Park, & Yoon, 2019). Perplexity is a statistical measure of how well a probability model predicts a sample. As applied to LDA, for a given value of *k*, you estimate the LDA model. Then given the theoretical word distributions represented by the topics, compare that to the actual topic mixtures, or distribution of words in your documents.

Perplexity is seen as a good measure of performance for LDA. Based on the research of Kannan, Mahalakshmi, Smitha, and Sendhilkumar (2018), The idea is that you keep a holdout sample, train your LDA on the rest of the data, then calculate the perplexity of the holdout. The perplexity could be given by the formula:

$$per(D_{test}) = exp\{-\frac{\sum_{d=1}^{M} \log p(\mathbf{w}_d)}{\sum_{d=1}^{M} N_d}\} \qquad \textit{Equation 4}$$

Here *M* is the number of documents (in the test sample, presumably), *wd* represents the words in document *d*, $N_d$ the number of words in document *d*.

To evaluate the model's perplexity, the researcher tested the data set with several k loops from 2-20. The perplexity score is then computed through R and is displayed in a graph.

The perplexity metric is a predictive one. It assesses a topic model's ability to predict a test set after having been trained on a training set. In practice, around 80% of a corpus may be set aside as a training set with the remaining 20% being a test set. To perform this evaluation, the researcher used the topic model library in R language particularly the perplexity() function as shown in Figure 5.

```
49  train = sample(rownames(dtm), nrow(dtm) * .80)
50  dtm_train = dtm[rownames(dtm) %in% train, ]
51  dtm_test = dtm[!rownames(dtm) %in% train, ]
52
53  m = LDA(dtm_train, method = "Gibbs", k = 2,  control = list(alpha = 0.01))
54
55
56
57  #PERPLEXITY
58  perplexity(m, dtm_test)
```

*Figure 5*. Code snippet of the perplexity function used by the researchers.

The division of the train set and the test set with the former being assigned 80% of the corpus while the latter being assigned the remaining 20% is shown in line 49 of Figure 5. However, the statistic is somewhat meaningless on its own. The benefit of this statistic comes in comparing perplexity across different models with a varying number of k. The model with the lowest perplexity is generally considered the "best". To do this, the researcher created a loop in R that iterates the number of k starting from the value of 2

up to 20 which is illustrated. Figure 6 shows the R loop function developed to perform perplexity on k values ranging from 2 to 20.

```
#PERPLEXITY
perplexity(m, dtm_test)
p = data.frame(k = c(2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20),
               perplexity = NA)

## loop over the values of k in data.frame p
for (i in 1:nrow(p)) {
  print(p$k[i])
  ## calculate perplexity for the given value of k
  m = LDA(dtm_train, method = "Gibbs", k = p$k[i],  control = list(alpha = 0.01))
  ## store result in our data.frame
  p$perplexity[i] = perplexity(m, dtm_test)
}
```

*Figure 6*. Code snippet of the loop function to determine the perplexity per k-value.

## Coherence

Probabilistic coherence measures how associated words are in a topic, controlling for statistical independence (Jones, 2021). A set of statements or facts is said to be coherent if they support each other. Thus, a coherent fact set can be interpreted in a context that covers all or most of the facts (Kapadia, 2019).

According to Jones (2021), a topic with the words {sport, sports, ball, fan, athlete} would look great if you look at correlation, without correcting for independence. But people know that it is a terrible topic because the words are so frequent in this corpus as to be meaningless. In other words, they are highly correlated with each other but they are statistically independent of each other.

For each pair of words {a,b} in the top M words in a topic, probabilistic coherence calculates P(b|a)–P(b), where {a} is more probable than {b} in the topic. To decide the optimum number of topics to be extracted using LDA, a topic coherence score is always used to measure how well the topics are extracted:

$$CoherenceScore = \sum_{i<j} score(w_i, w_j) \qquad \textit{Equation 5}$$

where $w_i$,$w_j$ are the top words of the topic.

Even though the perplexity metric is a natural choice for topic models from a technical standpoint, it does not provide good results particularly in human interpretation (Giri, 2021). This means that optimizing for perplexity may not yield human interpretable topics. This limitation of perplexity opened up an avenue for another evaluation tool that puts additionally puts into consideration the human judgment model which is Topic Coherence (Kapadia, 2019).

Additionally, Kapadia (2019) also explained that the concept of topic coherence combines several measures into a framework to evaluate the coherence between topics inferred by a model. Its measures score a single topic by measuring the degree of semantic similarity between high-scoring words in the topic. These measurements help distinguish between topics that are semantically interpretable topics and topics that are artifacts of statistical inference.

To get the coherence score of the test dataset, the researcher used the R package textmineR which has the function that can calculate the coherence based on the document term matrix. The same approach with visualizing perplexity was also used to get the coherence. The R script loops from 1 to 20 values for k and calculates the coherence score for each k value.

```
k_list <- seq(1, 20, by = 1)
model_dir <- paste0("models_", digest::digest(vocabulary, algo = "sha1"))
if (!dir.exists(model_dir)) dir.create(model_dir)
model_list <- TmParallelApply(X = k_list, FUN = function(k){
  filename = file.path(model_dir, paste0(k, "_topics.rda"))

  if (!file.exists(filename)) {
    m <- FitLdaModel(dtm = dtm, k = k, iterations = 2000)
    m$k <- k
    m$coherence <- CalcProbCoherence(phi = m$phi, dtm = dtm, M = 5)
    save(m, file = filename)
  } else {
    load(filename)
  }

  m
}, export=c("dtm", "model_dir")) # export only needed for Windows machines
```

*Figure 7*. R script code snippet for calculating the Coherence for k values 1-20.

## RESULTS AND DISCUSSIONS

To design and develop the system, the researchers first conducted an initial interview with the records officer of the Sangguniang Panlalawigan to initialize the first step of the SCRUM development model which was the product backlog. In this stage, features that the client wanted to find and suggested to be included were specified and were noted by the researcher. The researchers then created a product backlog file to monitor the status of each feature.

Based on the product backlog, sprint planning took place on whether which functions were short and long sprints. Relative features that belong to the same module are grouped into one sprint while also taking into consideration the priority level of each function with the highest priority level being the functions labeled as 'MUST'. Each sprint processed and finished was reviewed by the researcher then checked whether the developed module conformed to the specifics provided in the product backlog. The researcher then consulted with and showed the result of the development for each sprint to the Records Officer of the Sangguniang Panlalawigan to check whether they were satisfied with the outcome of the performed sprint.

The system was developed using PHP as the backend development language. HTML5 and CSS3 were used to create an aesthetically pleasing user interface. Bootstrap

was used by the researcher to ensure that the system is responsive and adjusts to different screen sizes. Javascript and Jquery were used to enable the system to load data without refreshing the parent page window. To store the data to be processed by the system, a MySQL database was employed.

The developed system can be subdivided into two web portals. The administrator and the public portal. The Administrator portal is mainly intended for the system administrators and board members of the Sangguniang Panlalawigan. Its main functions are 1) to manage and track legislative documents of the province of Laguna; 2) generate and export lists of documents based on the search parameter specified and 3) apply natural language processing such as sentiment analysis and topic modeling on collected data sets using the Twitter Scraper module.

**Administrator Portal**

Figure 8 illustrates the documents list page where users can preview a list of the documents stored within the system. Available data for preview on this page includes the origin, document creator, document category, document type, document status, date created, accessibility status, and the QR code of the document.

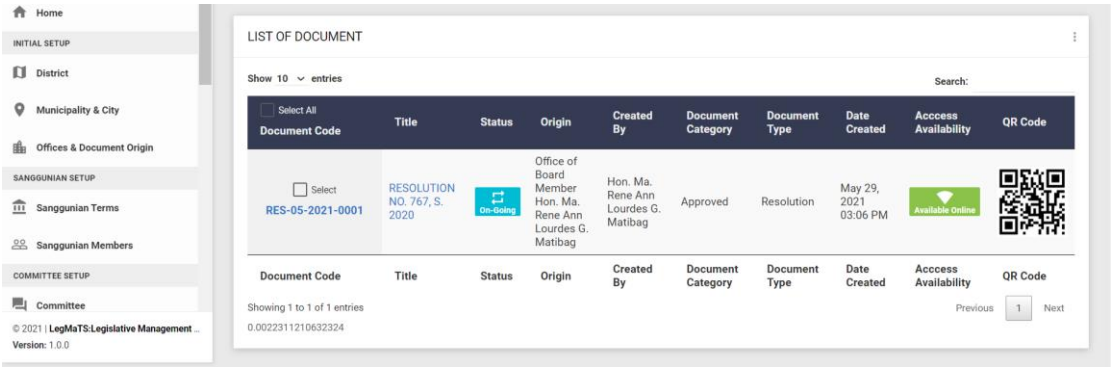


*Figure 8*. Documents list page.

Figure 9 shows the document management page which enables users to Add or Update existing document records in the system. The page allows users to encode the details of their documents. The page also includes features to enhance and optimize data encoding such as a What-you-see-is-what-you-get (WYSIWYG) editor which provides an editing toolbar for users as well as a system function that extracts texts from PDF files uploaded.

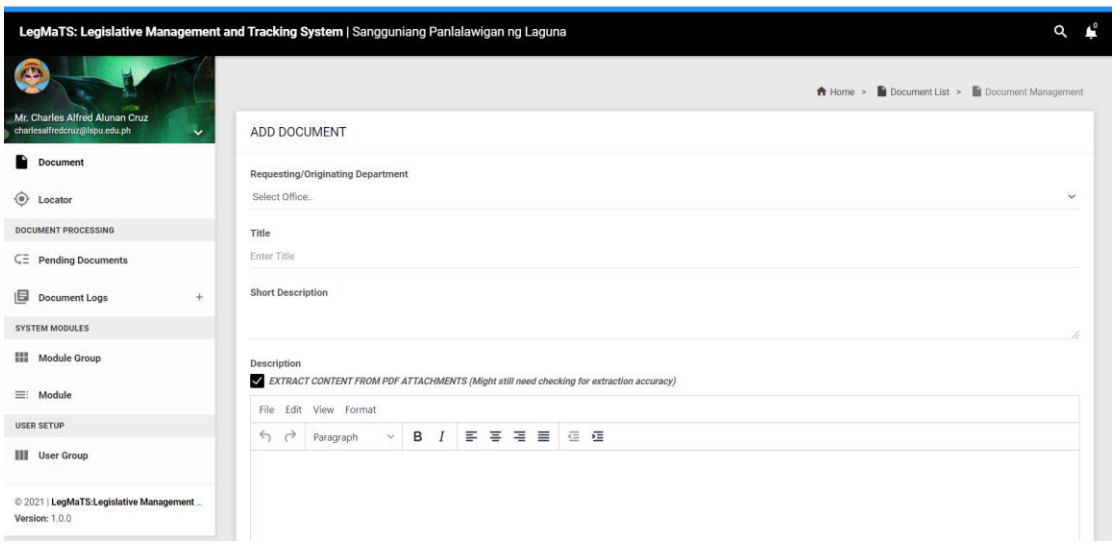


*Figure 9*. The add document management page with WYSIWYG editor and PDF to Text extractor.

Figure 10 shows the document trail page where users can preview document details, the document trail where the document went through, and the actual PDF attachment of the document. A Twitter developer account was registered for the system to be able to use the Twitter Search API. In doing so, the system was able to fetch public tweets based on the queries specified by the user.

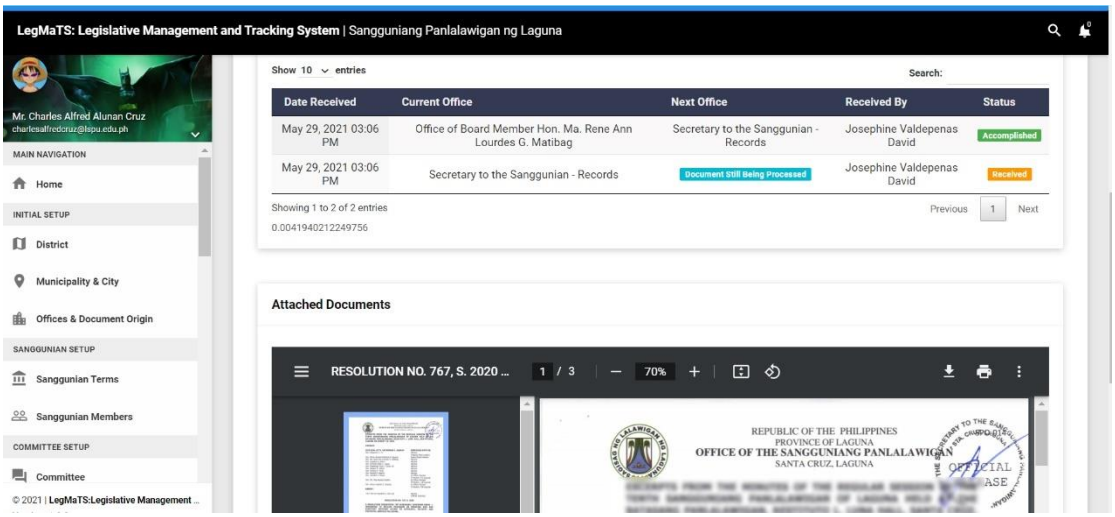


*Figure 10*. The document trail page.

Figure 11 shows the Twitter developer portal dashboard for the developed system which lets the researcher monitor the status of the system in terms of the Twitter search API.

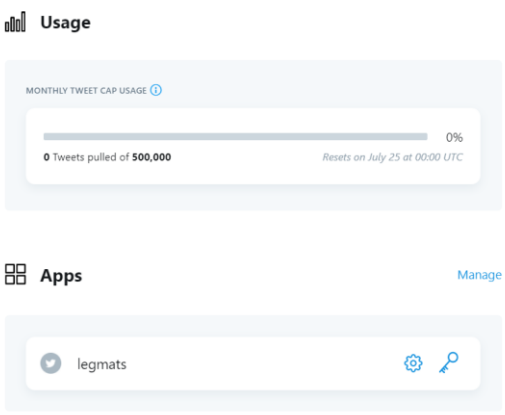


*Figure 11*. Twitter developer account dashboard for the developed system.

Figure 12 shows the Natural Language Processing reports page where users can create NLP transactions from the system and perform Topic Modeling from the scraped data such as collecting user data posts from Twitter as well as Sangguniang Panlalawigan documents stored within the system.

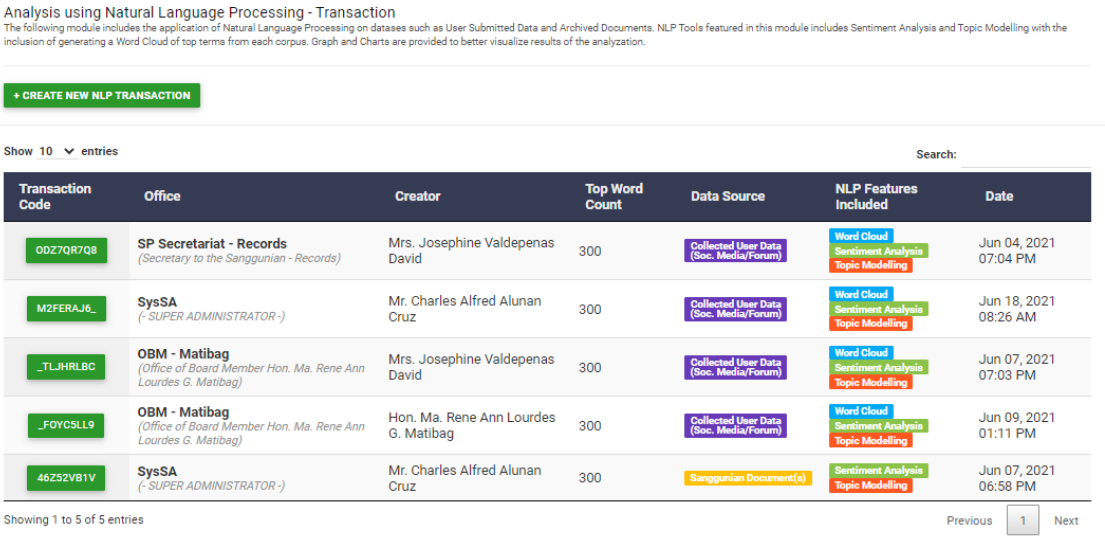


*Figure 12*. Twitter developer account dashboard for the developed system.

Figure 13 shows the Natural Language Processing transaction reports modal where the system asks the users which type of data they want to apply NLP analysis to. Options include the collected user data which tells the system to collect public tweets of users as well as posts to the public portal's forums page. Another option is Sanggunian Documents which tells the system to collect texts from stored documents to the system. Sentiment Analysis was only applied to collect user data while Topic Modeling was applied to all the data source options.

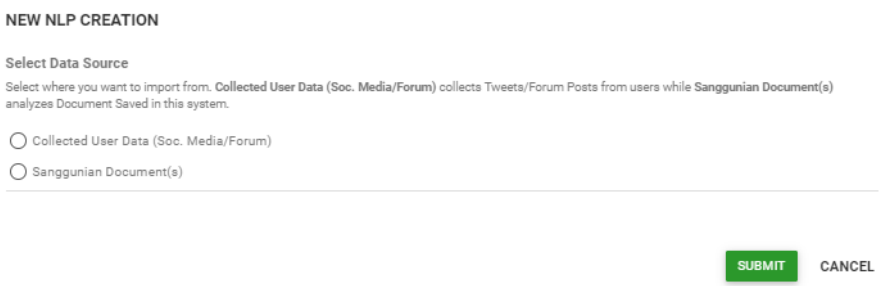


Figure 13. New NLP transaction modal.

To gather the data that was used for analysis, the researchers utilized the Twitter scraping tool developed. This system module adhered to Twitter's search API using a Twitter developer account. Public tweets were returned starting from June 21, 2021, to June 26, 2021. To specifically identify which tweets were posted within the vicinity of the province of Laguna, the coordinates (or geocode) of the Provincial Government of Laguna were also specified as a search parameter. A 25mi radius from the specified geocode was added as a search parameter to include the municipalities and cities that are nearby the specified center of the search.

Figure 14 shows a screenshot of the results from the Twitter scraper tool. The screenshot shows the username, the unprocessed and preprocessed version of the tweet as well as the translation of the tweet. The translation of the content can be achieved by either automated translation through Google's Translate API or by manually encoding the translation in the case of certain contents that are not accurately translated.

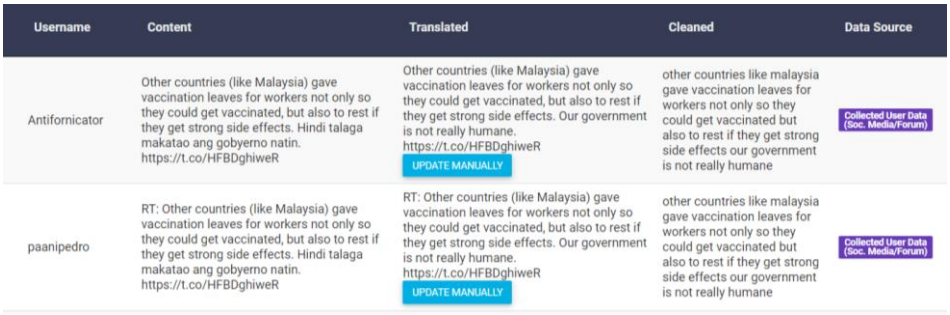


*Figure 14*. Screenshot of the dataset review page based on the gathered tweets through the Twitter Scraper Module.

**Observation-based evaluation**

The easiest way to evaluate a topic model is to look at the most probable words in the topic. To do this, the researcher listed the top 10 words and their frequency. Another way the researcher implemented this is to generate a word cloud to create a visually appealing way to observe the probable words in a topic.


Figure 15. Topic word cloud-based from the test data set.

To illustrate, Figure 15 is a word cloud based on topics modeled from the dataset collected. As you can see in the results of Word Clouds, words with higher frequency from the dataset appear larger as compared to words with lower frequency. Word clouds have emerged as a straightforward and visually appealing visualization method for text. They are used in various contexts as a means to provide an overview by distilling text down to those words that appear with the highest frequency (Heimerl, Lohmann, Lange & Ertl, 2014).

**Perplexity**

Figure 16 illustrates the result of the loop function in a line graph to visualize the perplexity of different k values from 2 to 20. Based on the graph, k with the value of 20 is the lowest perplexity.

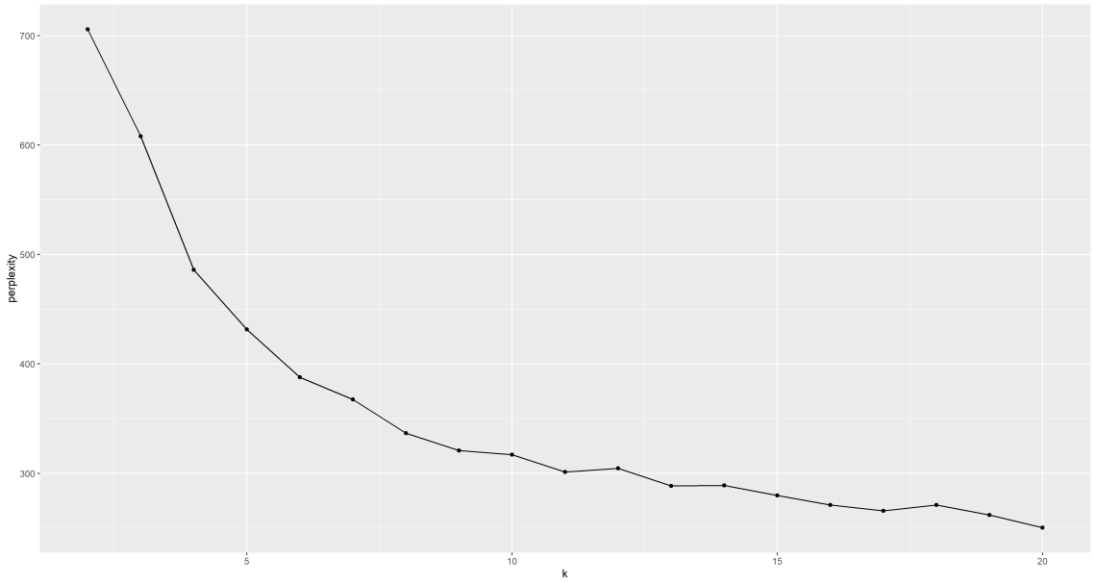


*Figure 16*. Perplexity graph of k values ranging from 1-20.

## Coherence

Figure 17 shows the coherence scores for various k values ranging from 1 to 20. Based on the graph, the k value with the highest coherence score is 6 which means that 6 is the best k value for the test dataset. The researcher then tested the model again using the system module developed. this time, based on the k value suggested by the coherence score evaluation. Moreover, optimizing for perplexity may not yield human interpretable topics. This limitation of perplexity measure served as a motivation for more work trying to model the human judgment, and thus Topic Coherence (Kapadia, 2019). Topic Coherence measures score a single topic by measuring the degree of semantic similarity between high scoring words in the topic. These measurements help distinguish between topics that are semantically interpretable topics and topics that are artifacts of statistical inference.

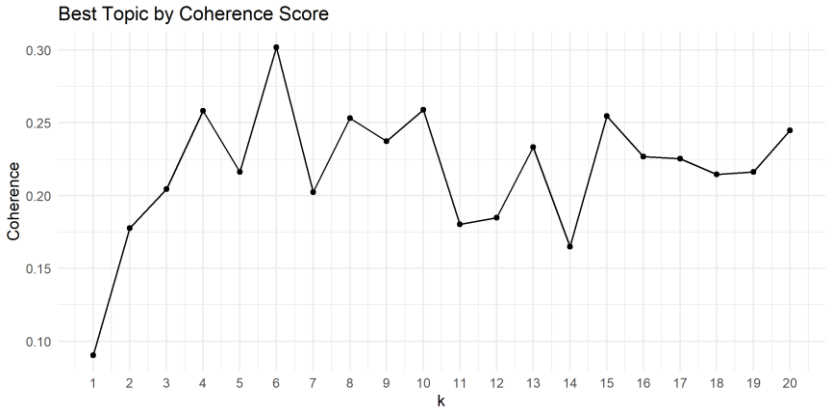


*Figure 17*. Topic coherence graph for values of k ranging from 1-20

## Application of Evaluation Results

LDA is being guided by two principles. It is that Every document is a mixture of topics and Every topic is a mixture of words. These are called Word-topic probabilities and Document-topic probabilities respectively. To gain insight, the researcher implemented these techniques to the dataset to generate the LDA result using the k value suggested by the coherence evaluation which is k=6.

. Figure 18 visualizes the 10 terms that are most common within each topic which applies the word-topic probabilities of an LDA model are the probabilities of observing each word in each topic of the LDA model. It is a V-by-K matrix, where K is the number of topics and V is the number of words in Vocabulary. The (v,k)th entry of Word-Topic probability corresponds to the probability of observing word v in topic k (Mathworks, 2021). Additionally, the visualization provides users an easier way to understand the topics that were extracted from the test dataset.

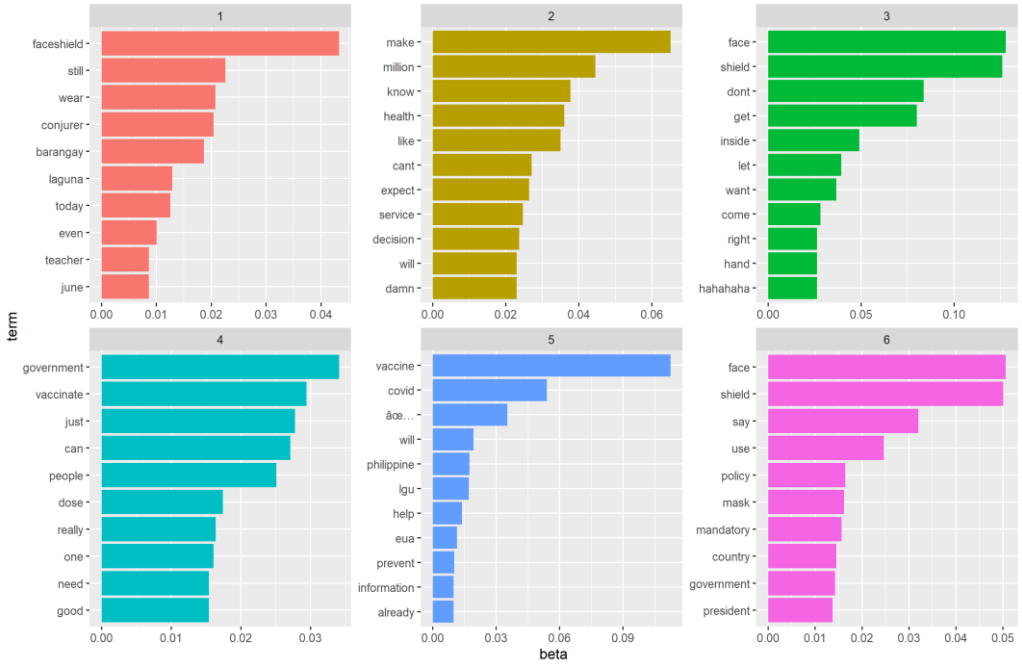


*Figure 18*. Word-topic probabilities visualization.

**Document-Topics probabilities**

Besides estimating each topic as a mixture of words, LDA also models each document as a mixture of topics. The researcher examined the er-document-per-topic probabilities, called γ ("gamma"), with the matrix = "gamma" argument to tidy() R package. Each of the values generated is an estimated proportion of words from that document that are generated from that topic. For example, based on Table 1, the model estimates that only about 19% of the words in document 6 were generated from topic 1.

Figure 19 illustrates the Top 6 topics by prevalence in the test dataset with the top words that contribute to each topic as a result of combining the beta and the gamma matrices. The graph was also arranged by the most coherent topic which is Topic 3.

**Visualization of Results**

The developed system visualizes the results as an interactive graph using the LDAvis function in R. LDAvis is designed to help users easily interpret and understand the topics generated from the topic model which has been fitted to text data from a corpus. It extracts the information generated based on the fitted LDA topic model and displays the result in an interactive termite-based visualization to provide users an easily understood representation (Sievert C. & Shirley K., 2014).

Table 1. Top 18 Document-topics probabilities

| # | DOCUMENT | TOPIC | GAMMA |
|---|---|---|---|
| 1 | 1 | 1 | 0.169697 |
| 2 | 2 | 1 | 0.135266 |
| 3 | 3 | 1 | 0.148148 |
| 4 | 4 | 1 | 0.153005 |
| 5 | 5 | 1 | 0.122549 |
| 6 | 6 | 1 | 0.187879 |
| 7 | 7 | 1 | 0.111111 |
| 8 | 8 | 1 | 0.112613 |
| 9 | 9 | 1 | 0.155251 |
| 10 | 10 | 1 | 0.182796 |
| 11 | 11 | 1 | 0.247126 |
| 12 | 12 | 1 | 0.194969 |
| 13 | 13 | 1 | 0.122549 |
| 14 | 14 | 1 | 0.111111 |
| 15 | 15 | 1 | 0.179894 |
| 16 | 16 | 1 | 0.22549 |
| 17 | 17 | 1 | 0.111111 |
| 18 | 18 | 1 | 0.190476 |

The beta and gamma matrices were combined to better understand the topic prevalence in the dataset and which words contribute to each topic.


*Figure 19*. Top 6 topics by prevalence in the test dataset with the top words that contribute to each topic.

Figure 20 illustrates the results of the LDA in an interactive manner using LDAvis. The left panel of the visualization presents a global view of the topic model while the right panel of the visualization depicts a horizontal bar chart whose bars represent the individual terms that are the most useful for interpreting the currently selected topic on the left. The left and right panels of the visualization are linked together. By selecting a topic on the left, It reveals the most useful terms on the right for interpreting the selected topic. Additionally, selecting a term on the right shows the conditional distribution over topics on the left for the selected term. This kind of linked selection allows users to examine a large number of topic-term relationships in a compact manner (Sievert & Shirley, 2014).

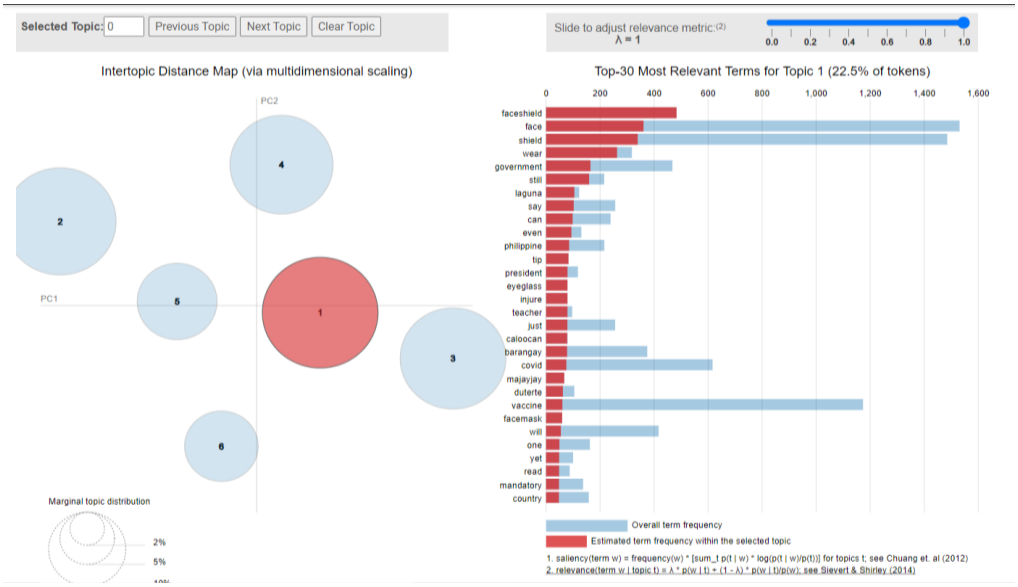


Figure 20. LDA results from visualization in the developed system using LDAvis.

Furthermore, the visualization of the topic model as a result of the system gives the Sangguniang Panlalawigan an overview as to what topics are most talked about and which words consist in each topic. It can be a tool/medium for legislators to identify the citizen's topic of concern within a specific time frame.

## CONCLUSIONS AND RECOMMENDATIONS

The study presents a system that allows legislators of the province of Laguna to collect public posts from the social networking site Twitter and used Latent Dirichlet Allocation (LDA) topic modeling to reveal abstract topics from archived documents and social media posts collected using Twitter. The system helps the administrators of Panlalawigan ng Laguna to easily identify the most frequent concerns and issues of the community that leads to formulating policies and ordinances appropriate for them. The users of the system can now easily manage, track and utilize the document tracking and assess the common topics in the collected documents through the system. A Latent

Dirichlet Allocation (LDA) model was successfully developed using R and interfacing the result in an interactive termite page to easily understand the result. Although the k number was a gut feeling, the model was also evaluated through the same collected data set from the Twitter Scraper tool using the Perplexity and Coherence scores and resulted in a possible optimal value for k which was 20 and 6 respectively. To provide the users with a cleaner and easily understandable output, an interactive graph that allows users to explore the LDA model generated by the system was provided through the topic modeling page module on the administrator portal. This was achieved using the LDAvis package in R. It is recommended to implement the developed system no less than 5 years to observe and identify additional improvements to the system. Additionally, It is also recommended that the system be tested on a much larger dataset to fully evaluate the performance of the LDA model. It is also recommended to improve the list of the stop words and noise removal feature to improve the result of the topic model. Additionally, it is also recommended to apply the Bi-Gram and Tri-gram approach to Topic Modeling to further enhance the words generated within each topic.

## IMPLICATIONS

Governments are now implementing the power of technology which help to communicate and interact with the citizen and their community. Through the integration of Information Communication and Technology (ICT), everyone can easily do their task such as managing and tracking documents, being involved in the different discussions and concerns as well as being empowered.

The system can be used to simply accomplish the document trail page where users can preview document details which is a big help for some transactions in the office. Moreover, administrators and personnel staff can identify the needs of the community, assess the thoughts and attitudes of the community towards the formulated programs through the application of NLP transactions from the system, and perform Topic Modeling from the scraped data such as collecting user data posts from Twitter as well as Sangguniang Panlalawigan documents stored within the system.

The implementation of different visualization techniques like LDAviz and Word Cloud in the system helps to facilitate the to provide an impression by extracting text down to words that are the most concerns or issues of the citizen, and topics that can be a basis of crafting programs and priorities of the government officials in taking actions to the citizen concerns.

## REFERENCES


Alguliyev, R. M., Aliguliyev, R. M., & Niftaliyeva, G. Y. (2019). Extracting social networks from e-government by sentiment analysis of users' comments. *Electronic Government, an International Journal*, *15*(1), 91-106.

Beck M. (2020). *How to scrape tweets from Twitter*. Retrieved on 07/15/2021, from https://towardsdatascience.com/how-to-scrape-tweets-from-twitter-59287e20f0f1

Buccafurri, F., Fotia, L., & Lax, G. (2015). A privacy-preserving e-participation framework allowing citizen opinion analysis. *Electronic Government, an International Journal, 11*(3), 185-206.

Ganesan, K. (n.d). *What are N-grams?*. Retrieved January 30, 2022, from https://kavita-ganesan.com/what-are-n-grams/#.YeJnxNFBzIU. January 14, 2022.

Giri (2021).*Topic model evaluation.* Retrieved on 01/13/2022, Retrieved from https://highdemandskills.com/topic-model-evaluation/

Heimerl, F., Lohmann, S., Lange, S., & Ertl, T. (2014). Word cloud explorer: Text analytics based on word clouds. In *2014 47th Hawaii International Conference on System Sciences* (pp. 1833-1842). IEEE.

Hubert, R. B., Estevez, E., Maguitman, A., & Janowski, T. (2018). Examining government-citizen interactions on Twitter using visual and sentiment analysis. In *Proceedings of the 19th annual international conference on digital government research: governance in the data age* (pp. 1-10).

Jelonek, D., Stępniak, C., Turek, T., & Ziora, L. (2020). Planning cities development directions with the application of sentiment analysis. *Prague Economic Papers*, *2020*(3), 274-290.

Jeong, Y., Park, I., & Yoon, B. (2019). Identifying emerging Research and Business Development (R&BD) areas based on topic modeling and visualization with intellectual property right data. *Technological Forecasting and Social Change, 146*, 655-672.

Jones T.W. (2014).*Topic modeling*. Retrieved on 07/15/2021, from https://cran.r-project.org/web/packages/textmineR/vignettes/c_topic_modeling.html

Jose B. K. (2021).*Data Preprocessing | Natural Language Processing.* Retrieved on 01/25/2021, from https://basilkjose.medium.com/data-preprocessing-natural-language-competition-processing-dcbbf9d014e8

Kannan, M. S., Mahalakshmi, G. S., Smitha, E. S., & Sendhilkumar, S. (2018). A word embedding model for topic recommendation. In *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)* (pp. 826-830). IEEE.

Kapadia S. (2019). *Evaluate Topic Models: Latent Dirichlet Allocation (LDA)*. Retrieved on 01/13/2022, Retrieved from https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0

Lin, F. R., Chou, S. Y., Liao, D., & Hao, D. (2015, January). Automatic content analysis of legislative documents by text mining techniques. In *2015 48th Hawaii International Conference on System Sciences* (pp. 2199-2208). IEEE.

Mathworks. (2021). *LDA model*. Retrieved on 01/25/2022, from https://fr.mathworks.com/help//textanalytics/ref/ldamodel.html

Pablo M. (2018). *Internet Inaccessibility Plagues "Social Media Capital of the World"*. Retrieved on 01/13/2022, Retrieved from https://asiafoundation.org/2018/10/24/internet-inaccessibility-plagues-social-media-capital-of-the-world

Pascual F. (2019).*Topic Modeling: An Introduction*. Retrieved on 07/15/2021, Retrieved from https://monkeylearn.com/blog/introduction-to-topic-modeling/

Porio, E. (2017). Citizen participation and decentralization in the Philippines. In *Citizenship and democratization in Southeast Asia* (pp. 29-50). Brill.

Porteous, I., Newman, D., Ihler, A., Asuncion, A., Smyth, P., & Welling, M. (2008, August). Fast collapsed Gibbs sampling for latent Dirichlet allocation. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 569-577).

Rabindranath G. (2020). *Topic Model Evaluation.* Retrieved on 07/15/2021, Retrieved from https://towardsdatascience.com/topic-model-evaluation-3c43e2308526#2932

Salleh, S. F., Ujir, H., Sapawi, R., & Hashim, H. F. (2020). Accreditation Document Tracking System Using Scrum Approach. *International Journal of Evaluation and Research in Education, 9*(1), 153-161.

Senthilkumar, R., RubanRaja, B., & Monisha, -. (2021). Brand Positioning and Segmentation of Sneakers through Multi-Dimensional Customer Experience Analysis. *Journal of Scientific Research, 13*(2), 335–345. https://doi.org/10.3329/jsr.v13i2.47841

Sievert, C., & Shirley, K. (2014, June). LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces* (pp. 63-70).

Subeno, B., & Kusumaningrum, R. (2018). Optimization towards Latent Dirichlet Allocation: Its Topic Number and Collapsed Gibbs Sampling Inference Process. *International Journal of Electrical & Computer Engineering (2088-8708), 8*(5),3204-3213.

Wowchemy. (2021). *Topic modeling*. Retrieved on 01/25/2022, Retrieved from https://cfss.uchicago.edu/notes/topic-modeling/

Yusingco, M. H. (2020, October), "Social media and democracy in the Philippines", Retrieved on 07/15/2021, Retrieved from https://blogs.griffith.edu.au/asiainsights/social-media-and-democracy-in-the-philippines/

Zaidi, S. F. H., & Qteishat, M. K. (2012). Assessing e-government service delivery (government to citizen). *International journal of ebusiness and egovernment studies, 4*(1), 45-54.