# Email Clustering & Generating Email Templates Based on Their Topics

Fatih Coşkun
Research & Development Center, adesso Turkey, Istanbul, Turkey
Fatih.Coskun@adesso.com.tr

Cengiz Gezer
Research & Development Center, adesso Turkey, Istanbul, Turkey
Cengiz.Gezer@adesso.com.tr

V. Çağri Güngör
Department of Computer Engineering, Abdullah Gul University, Kayseri, Turkey
Cagri.Gungor@agu.edu.tr

## ABSTRACT

Email templates have a significant impact on users in terms of productivity. Using an email template that is produced successfully is going to transfer the main information with a considerable impression. While the previous studies were focused on the email generation by text-differences in the content of the emails, generated templates based on email topics can provide better productivity for the companies. This article proposes a system, in which user emails are clustered according to the topics of the emails, and introduces an email template generation system that utilizes the sample emails belonging to the formed email clusters. For this purpose, the Enron email dataset has been used and the performance of different text preprocessing and topic modeling algorithms, such as DMM, GPU-DMM, GPU-PDMM, LF-DMM, LDA, LF-LDA, BTM, WNTM, PTM, SATM, have been investigated and compared to determine the most efficient one. After obtaining the email topics, the system shows the examples of the emails representing the selected topics and enables the authorized users to create templates that generalize these topics.

## CCS CONCEPTS

• **Computing methodologies** → Machine learning.

## KEYWORDS

Topic modeling, Email Clustering, Template Generation, Short Text Topic Modeling, Effective Email Communication

**ACM Reference Format:**
Fatih Coşkun, Cengiz Gezer, and V. Çağri Güngör. 2021. Email Clustering & Generating Email Templates Based on Their Topics. In *2021 the 5th International Conference on Information System and Data Mining (ICISDM 2021), May 27–29, 2021, Silicon Valley, CA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3471287.3471298

## 1 INTRODUCTION

Email templates have a significant impact on users in terms of productivity. Using an email template that is produced successfully is going to transfer the main information with a considerable impression. While the previous studies were focused on the email generation by text-differences in the content of the emails, generated templates based on email topics can provide better productivity for the companies.

Gathering the emails that have the same topics during the creation of email templates will allow users to obtain email templates that generalize the topics determined for all users using the email system. Rather than identifying email templates, identifying the topics of existing emails, aiming to turn these clusters into email templates can strengthen the context that organizations communicate with by email.

Emails that have similar content and are related to a limited number of topics, especially according to the field of study in corporations, are mostly used (billing information, meeting organization, etc.). These emails are rewritten by users every time. This situation is open to both probable errors and time loss due to the increase of mutual messaging and reasons, such as missing information.

Ready-made templates for messaging in the limited number and content mentioned above can be presented by institutions or software providers. However, since these ready-made templates change from institution to institution or evolve over time, they need to be updated manually. In the current technique, solutions addressing these difficulties are not offered. Also, rarely used emails may become widespread in the company over time and it is another manual workload to recognize these and provide templates for these emails.

This article proposes a system, in which user emails are clustered according to the topics of the emails and introduces an email template generation system that utilizes the sample emails belonging to the formed email clusters. For this purpose, the Enron email dataset has been used and the performance of different feature selection and topic modeling algorithms, such as DMM [1], GPU-DMM [2], GPU-PDMM [3], LF-DMM [4], LDA [5], LF-LDA [4], BTM [6], WNTM [7], PTM [8] and SATM [9], have been investigated and compared to determine the most efficient one. After obtaining the email topics, the system shows the examples of the emails representing the selected topics and enables the authorized users to create templates that generalize these topics. In general, the main contributions of this paper are as follows:

- Determining the most appropriate clustering model by comparing the topic modeling algorithms that can be applied to the email datasets.
- To reveal that improvements in topic modeling algorithms, provide a noticeable improvement in results

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
*ICISDM 2021, May 27–29, 2021, Silicon Valley, CA, USA*
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8954-9/21/05…$15.00
https://doi.org/10.1145/3471287.3471298

**Table 1: Topic modeling methods on emails**

| Reference | Year | Dataset | Method |
|---|---|---|---|
| Dredze et al. [10] | 2008 | Enron | LDA & LSA-based |
| Ayodele and Zhou [11] | 2008 | Not Specified | Evolving Clustering Method |
| Ayodele et al. [12] | 2009 | Enron (4000 emails) | Email Evolving Clustering Method |
| Joty et al. [13] | 2010 | BC3 Corpus | LCSeg & LDA |
| Yang et al. [14] | 2010 | 20NewsGroup | K-means clustering |
| Pan et al. [15] | 2013 | Private 7000+ & New York Times Articles (13000+) | CITCC |
| Hong and Moh [16] | 2015 | Enron | Developed LDA |

**Table 2: Topic modeling methods for short texts**

| Reference | Year | Algorithm |
|---|---|---|
| Yin and Wang | 2014 | Dirichlet Multinomial Mixture (DMM) |
| Yan et al. | 2013 | Biterm Topic Model (BTM) |
| Zuo et al. | 2015 | Word Network Topic Model (WNTM) |
| Zuo et al. | 2016 | Pseudo-Document-Based Topic Model (PTM) |
| Quan et al. | 2015 | Self-Aggregation-Based Topic Model (SATM) |
| Li et al. | 2016 | Generalized Pólya Urn (GPU) based Dirichlet Multinomial Mixture Model (GPU-DMM) |
| Li et al. | 2017 | Generalized Pólya Urn (GPU) based Poisson-based Dirichlet Multinomial Mixture Model (GPU-PDMM) |
| Nguyen et al. | 2015 | Latent Feature Model with DMM (LF-DMM) |
| Blei et al. | 2003 | Latent Dirichlet Allocation (LDA) |
| Nguyen et al. | 2015 | Latent Feature Model with LDA (LF-LDA) |

- To introduce the topic-based template creation system from emails

To the best of our knowledge, this is the first study focusing on the effective clustering of emails according to their topics and providing efficient and detailed email templates with the help of a supervisor according to the email written by users who use the email service.

This paper is organized as follows. In section 2, the literature survey about the topic modeling and template generation is presented. In section 3, the dataset, its source, and the preprocessing operations applied to the dataset are explained. In section 4, the parameters of the topic modeling techniques and feature selection methods are stated. In section 5, the results of the topic modeling algorithms and feature selection methods are discussed, and the way of implementation of the proposed method into an email service is described.

## 2 LITERATURE SURVEY

When the studies on email were examined, it was found that the most used email dataset suitable for topic modeling was the Enron email dataset and it was chosen as the email dataset to be used throughout this study.

There are several ways to extract topics from a dataset which are explained in the literature. However, because the dataset includes only emails, we have focused on topic modeling methods for emails and short texts. The studies on the topic modeling methods for emails can be seen in Table 1. As a result of examining the studies mentioned in Table 1 and the outputs of these studies, using LDA is a plausible reason for topic modeling.

Secondly, a literature survey about topic modeling techniques for short texts is performed for the reason that emails are also short texts. In recent years, it has been observed that studies on short text topic modeling have increased. While short text topic modeling has been applied to datasets such as Twitter or Google News data, there is a lack of studies in the literature that apply short text topic modeling to the emails by using the algorithms shared in recent years. For this purpose, using the STTM [17] (A Library of Short Text Topic Modeling), which contains the implementation of the algorithms in Table 2 and shared with open-source code, the results of the algorithms are obtained and compared using email data.

Furthermore, MALLET [18] has also been used for the study because of the advantages it provides in hyper-parameter optimizations and multithreaded operations. The main reason for using MALLET was to determine the number of the clusters (topics) with a model whose hyper-parameters were automatically adjusted.

As a third step, a literature survey for template generation is performed. However, when the methods used to obtain the templates from the emails were examined in the literature as it can be seen in Table 3, it is observed that the emails are not clustered by their topics, but by their structural differences of the texts. After the clustering operation, emails are transferred into the templates by removing the variable values. For example, Zhang et al. [19],

**Table 3: Template generation methods**

| Reference | Year | Method |
|---|---|---|
| Zhang et al. | 2015 | Text Similarity |
| Proskurnia et al. | 2017 | Text Similarity |
| Whittaker et al. | 2019 | Text Similarity |
| **Proposed Solution** | **2021** | **Topic Similarity** |

Proskurnia et al. [20], and Whittaker et al. [21] have focused on determining the variable data contained in or to be found in email and using them with templates. These studies have the potential to be used in the future stages of our work. However, after the emails are separated to create templates according to their topics, the email templates to be created will be more useful both in terms of number, and to create generalizing and more effective emails. For example, there can be several ways to organize a meeting over the emails by using completely different words but creating one template impressively instead of several templates will be more effective in terms of productivity.

## 3 DATASET AND TEXT PREPROCESSING

The source of the dataset is Enron Email Dataset, which has over 500,000 emails from Enron Corporation employees. With the investigation of Enron's collapse, it was publicly shared by the Federal Energy Regulatory Commission. The version shared at Carnegie Mellon University in 2015 was used for this study. The emails in the dataset include both the emails created by the real users and the emails created automatically in the HTML structure by the companies. Most of the real user emails are stored as email threading after the operations of forwarding and replying inside the database.

### 3.1 Dataset Features

The fields of an email in the dataset can be listed as the following: Message-ID, Date, From, To, Subject, Mime-Version, Content-Type, Content-Transfer-Encoding, X-From, X-To, X-cc, X-bcc, X-Folder, X-Origin, X-Filenames, Message Content.

### 3.2 Text Preprocessing Methods

An email in the dataset contains some text fields that are not related to the topics but related to the sender and receiver information. Therefore, the emails should be cleaned by several methods. In this way, the topic modeling algorithms will not try to handle the strings which are not related to the topics.

The fields of subject and message-content are the only ones that are directly related to the topic of the emails. Therefore, the other fields have been removed from the emails. Additionally, if the message has been forwarded or replied to, the message-content also contains some of the fields that are mentioned above. To increase the performance of the topic modeling algorithms, these fields have also been removed by using regex matching tools.

Since email addresses are considered to not affect the topic, while preprocessing is done on the dataset, email addresses are also removed from the message content. In this way, basic data to determine the topic of an email were obtained. After the concatenation of the subject and cleaned message-content, the following most-known text preprocessing operations have been performed to the whole dataset:

- Removing the parts which do not represent the topic from Dataset
- Lowercasing
- Getting the pure text from HTML
- Spell Checking and Autocorrection
- Removing Punctuation Marks & Stop words
- Filtering words by their frequency
- Stemming & Lemmatization
- Bigrams and Trigrams

After the mentioned text preprocessing methods, an email in the dataset seems like Figure 1.

## 4 METHODOLOGY

Topic Modeling is one of the most interesting areas of Machine Learning and Natural Language Processing. It helps to learn a general overview of the documents by examining the words in those documents. If the text is very long and complex data, it is very difficult to read and classify all these data. However, topic modeling methods are used to understand the main topics in a dataset. With this method, it is easier to identify the most valuable information even if the available data is large enough for a human. After the text preprocessing methods, various topic modeling algorithms were used to determine the most successful topic modeling algorithm that can work in a data set such as the Enron email dataset. STMM library, which is an open-source library and has an article with detailed explanations of methods for comparison of algorithms, was prepared according to our dataset.

### 4.1 Selected Hyper-parameters for Topic Modeling Techniques

MALLET is a tool created using JAVA API, which provides text-based machine learning applications such as topic modeling, natural language processing, text clustering, text classification. Because MALLET is an effective tool and it uses LDA, it is applied to the model to increase the performance of the clustering. Various hyper-parameters are set by using MALLET. One of the most significant hyper-parameters that cannot be selected directly by using MALLET is the number of topics. To decide the number of topics, the coherence values of the topics are used. The optimum number for the topics should have the maximum coherence value in a range. For the 100,000 emails (as a randomly selected sample of Enron Dataset), 26 is selected as the optimum number of the topics since it gave the maximum coherence value. In addition to the number of

system close may select review suggest list quickly possible feedback process can begin review approve supervisor access pep question regard contact help

**Figure 1: An example of an email in the Enron dataset after operations of text preprocessing**

topics parameter, each algorithm expects other parameters unique to itself. Other parameters used are listed below:

- **DMM, BTM & WNTM:** Alpha: 0.1, Beta: 0.01, Gibbs sampling iterations: 1000
- **LDA:** Alpha: 0.05, Beta: 0.01, Gibbs sampling iterations: 1000
- **LF-DMM & LF-LDA**: Alpha: 0.1, Beta: 0.01, Lambda: 0.6, Gibbs sampling iteration: 1000
- **PTM**: Alpha: 0.1, Beta: 0.01, Gibbs sampling iterations: 1000, Number of long pseudo-documents: 300
- **SATM:** Alpha: 0.1, Beta: 0.01, Gibbs sampling iterations: 1000, Number of long pseudo-documents: 300, Threshold: 0.001
- **GPU-DMM:** Alpha: 0.1, Beta: 0.01, Threshold: 0.1, Weight: 0.7, Filter Size: 20, Gibbs sampling iteration: 1000
- **GPU-PDMM:** Alpha: 0.1, Beta: 0.01, Lambda: 0.6, Threshold: 0.1, Weight: 0.7, Filter Size: 20, Gibbs sampling iteration: 100

### 4.2 Additional Dataset for Algorithms and Evaluation

For the GPU-DMM, GPU-PDMM, LF-DMM, and LF-LDA models to work, pre-trained word2vec [22] vectors had to be given to algorithms. For this reason, Google News data containing more than 100 billion words was downloaded and used via Google Code [23]. The main reason for using this dataset was the high number of words it contained and the ability to use the semantic relationship between the words at a remarkable level.

In addition to the Google News pre-trained data, Wikipedia [24] data to be used in the calculation of coherence values for the evaluation and comparison of models was downloaded and processed and included in the system by using the methods shared in the STTM library.

### 4.3 Feature Selection

The email dataset contains a lot of words for the topic modeling algorithms, although it has been filtered through various preprocessing algorithms to basic words and phrases that represent the topic. This situation reduces the working speed of the algorithms considerably. Because these methods will be used in companies, and the number of emails received and sent daily within a company is very high, it is a necessity to make these algorithms faster with some techniques. The most basic, effective, and preferred methods for reducing the required processing load without affecting the topic modeling algorithms are feature selection methods. Feature selection is used to detect and remove words in the dataset that may not have a positive effect on subject detection or even have a negative effect. By extracting these words, a "dictionary" will be created that can be used for company-specific topic modeling. This "dictionary", which is specific to the company and e-mail data, can be used as a filtering purpose when it is desired to make topic modeling with newly added emails in the future after its preparation.

To apply feature selection to the data, a 2-step process has been carried out:

1. The data has been labeled according to the topic numbers after the topic modeling algorithm has been selected and applied to the data.
2. By using labeled data, 3 feature selection method were compared (Chi-square Test, Information Gain, Mutual Information)

## 5 RESULTS & DISCUSSION

### 5.1 Evaluation of Topic Modeling Algorithms

Although different methods can be used to evaluate topic modeling algorithms, email data should be evaluated according to coherence values since they are not labeled. To calculate the topic coherence valuations, evaluation models over STTM were run using the previously downloaded Wikipedia file. After evaluation for each algorithm, coherence values were obtained. The comparison graph of the coherence values and the algorithms is in Figure 2.

According to these results, LF-LDA and LDA were very close to each other and gave higher coherence results than others. Therefore, it was decided to use LF-LDA as the topic modeling algorithm to be used for emails.

With this study, it is also shown that other algorithms created with improvements in the existing algorithms give higher coherences as expected. As can be seen from the graph, LF-LDA has higher coherence values than LDA; GPU-DMM, GPU-PDMM, and LF-DMM have higher coherence values than DMM.

### 5.2 Applying Feature Selection Algorithms

Feature selection algorithms were run with data labeled according to the results of the LF-LDA topic modeling. To select the feature selection method, the results returned by the feature selection methods were used. According to the ranking of the features returned by each feature selection algorithm, the most effective 5000 words were selected and words that are not one of these 5000 words were extracted from the dataset. After this filtering process, the algorithms were compared by using LF-LDA again. The reason for using LF-LDA in comparing algorithms is that LF-LDA topic modeling gives a higher coherence value than other algorithms on the dataset without feature selection.

According to the results shown in Table 4, Information Gain gave better results than other feature selection methods, although there were not very high differences between the coherence values. For this reason, the dataset filtered with the most effective 5000 words according to Information Gain and the topic modeling techniques applied on the filtered dataset.

As it can be seen in Figure 3, the results obtained when using the dataset with feature selection applied, LF-LDA and LDA gave higher consistency results than others, as before. Also, these results, it is supporting the situation when feature selection was not applied;

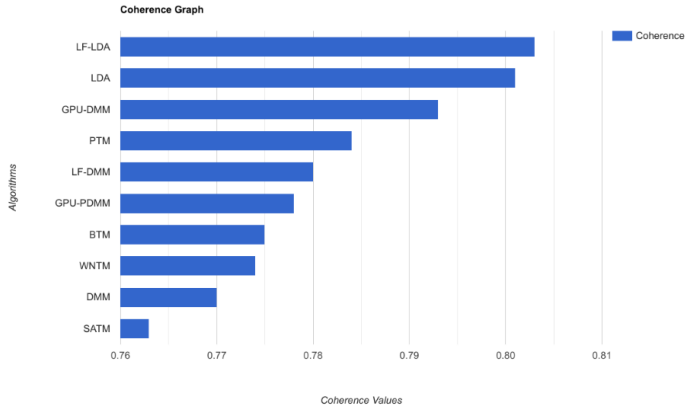


**Figure 2: Coherence result of topic modeling algorithms**

**Table 4: Comparison of coherence values with feature selection algorithms on the results of LF-LDA**

| Algorithm | Without Feature Selection | Information Gain | Chi-Square Test | Mutual Information |
|---|---|---|---|---|
| LF-LDA | 0.803 | 0.811 | 0.807 | 0.800 |

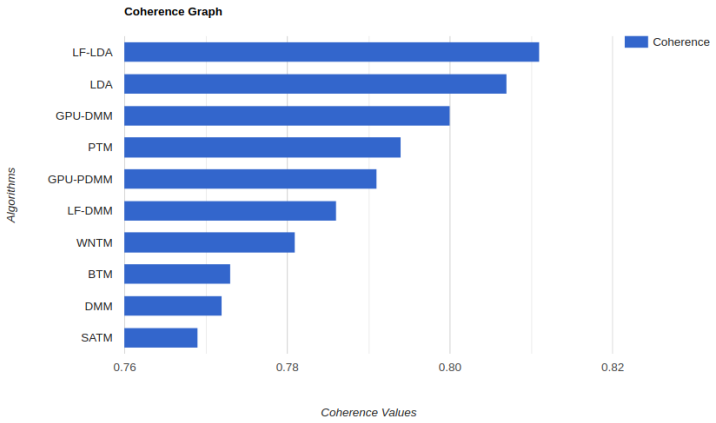


**Figure 3: Coherence result of topic modeling algorithms after feature selection**

LF-LDA has higher coherence values than LDA; GPU-DMM, GPU-PDMM, and LF-DMM have higher coherence values than DMM.

In addition, one of the best benefits of applying information gain is that it significantly reduces the required processing load. While the number of unique words in the dataset was over 14 thousand before the feature selection, this word count became 5 thousand. After this process, the running speed of the algorithms has speeded up.

## 5.3 Constituted Clusters

LF-LDA topic modeling was applied to the emails using the determined number of topics as a parameter. Some examples of topics can be seen in Table 5. The model can show the most used topics, the list of emails of the selected topic, and how much they belong to the detected topics of these emails.

**Table 5: Some of the topic words as the result of topic modeling**

| Topic 1 | Topic 2 | Topic 6 | Topic 9 | Topic 11 | Topic 15 | Topic 16 |
|---|---|---|---|---|---|---|
| agreement | travel | email | employee | schedule | meeting | access |
| attach | offer | message | program | time | discuss | service |
| file | day | information | attend | report | issue | image |
| comment | special | receive | provide | find | meet | click |
| document | purchase | address | follow | pm | work | link |
| draft | ticket | send | conference | datum | plan | url |
| review | visit | copy | request | system | regard | free |
| send | rate | contact | event | position | forward | online |
| form | fare | intend | process | hour | project | information |
| letter | ca | error | presentation | trade | discussion | internet |
| request | great | confidential | list | start | model | page |
| regard | include | offer | feedback | final | group | company |
| copy | night | delete | member | test | point | system |
| version | save | individual | management | table | suggest | site |
| sign | hotel | person | session | day | give | web |

## 5.4 Template Generation System

The compared algorithms are DMM, GPU-DMM, GPU-PDMM, LF-DMM, LDA, LF-LDA, BTM, WNTM, and PTM. We used the coherence values to evaluate the compared algorithms. Coherence values were calculated by using Wikipedia data and the results were between 0,76 and 0,81. LF-LDA and LDA models gave the highest coherence values and the LF-LDA algorithm was determined as the method to be used.

After comparing 3 feature selection algorithms (Information Gain, Mutual Information, and Chi-square Test), it is decided to use Information Gain as a feature selection technique due to higher coherence on results. This operation increased the coherence of LF-LDA from 0.803 to 0811 and the running time of the algorithms decreased considerably.

Besides, the LF-LDA algorithm, which is created with various improvements on LDA, gave a higher coherence result than LDA. Likewise, GPU-DMM, GPU-PDMM, and LF-DMM algorithms have been developed and have given higher coherence values than DMM.

After obtaining the topics using the emails, the system shows the examples of the emails that represent the selected topics, allowing the authorized users to create templates that generalize these topics. When this system is used for email templates within an organization, the communication quality of the company employees and the transmitted data can be effectively provided to the other party. The Email template creation phase can be automatized which is targeted as future work for each topic in the email database.

When the developed system is integrated into the email control panel, the system will begin to identify the topic by analyzing the data. At certain intervals, this system will show the created topics to the authorized user and will enable the authorized user to create effective and generalizing templates by using the emails that represent the topic. Employees/users will be able to see topic-based email templates created using a pre-trained model in their email systems and will be able to write effective emails by choosing the most appropriate one for them. In addition, when writing an email, the previously developed topic detection model determines the topic of the text, and possible email templates suitable for the topic are presented to the user. The steps to be taken for the integration of the email template generation module into the email system are described below.

*5.4.1 Steps to Be Taken by Supervisor.* As it can be seen in Figure 4, the following steps should be taken on the supervisor side.

1. In the study, corporate email databases are used as a dataset and newly written email is added to the dataset.
2. Data are filtered into the form of written sentences and text preprocessing algorithms are used to filter out words that are not important for topic modeling.
3. With the topic modeling techniques (LF-LDA, LDA, etc.), the data is divided into appropriate clusters according to the topics.
4. The keywords representing the cluster and their weights were determined statistically.
5. By comparing the keywords and each email data, the most obvious emails belonging to the clusters were revealed.
6. The model used in topic detection was recorded to identify the topics of new texts.
7. After the emails are clustered on the appropriate number of topics and the keywords of the topics are determined, template suggestions are created.
8. It is then displayed through a graphical interface so that an administrator (supervisor) can approve or change the templates.
9. The manager can open the templates he/she deems appropriate for general/private use within/outside the company.

*5.4.2 Steps to Be Taken on Email service.* As it can be seen in Figure 5, the following functionalities should be integrated into the email service.

1. When users who use the email service start writing an email, the module prepared for suggesting a template becomes active.

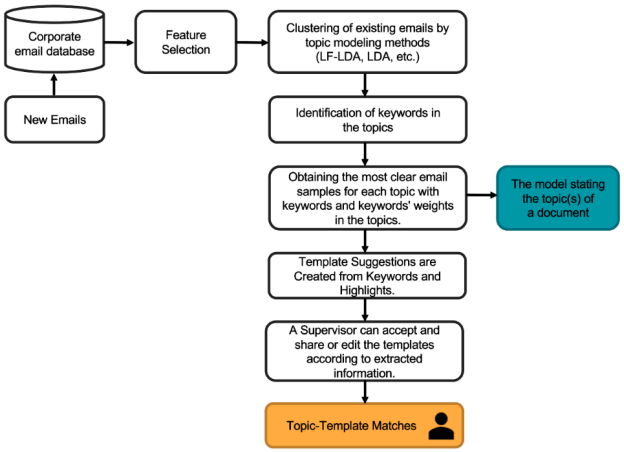


**Figure 4: Obtaining model and templates by automatically clustering existing emails**

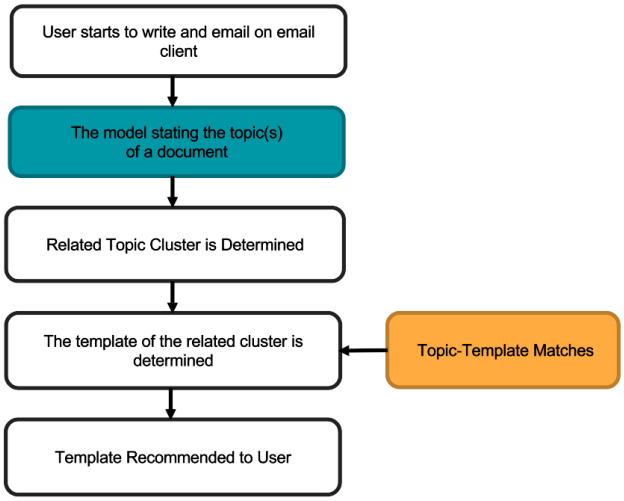


**Figure 5: Template suggestion model based on the email content by using a previously trained model**

2. The topic of the new email is determined while writing the email by the pre-trained topic clustering model.
3. Templates of the determined cluster are determined.
4. The specified templates are recommended to the user.

## 6 CONCLUSION

Email templates lead users to effectively improve their communication channels. Accurate and careful preparation of templates is critical for emails to be sent to the target audiences. Contrary to the systems created by making use of the written differences of the emails made earlier, in this article, we present a system that allows

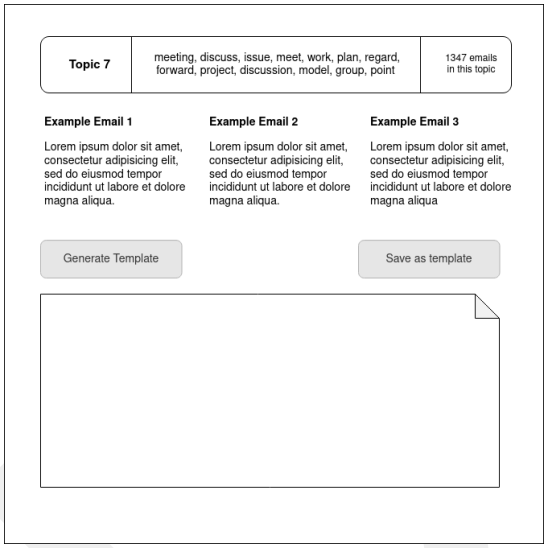


**Figure 6: The view of the admin control panel**

them to create specific templates for these categories by converting emails with the same topics but different contents in the categories.

In cases where manual methods are insufficient to analyze the data, topics that indicate importance in the texts can be determined by using methods such as topic modeling. In this way, generalizations can be made about the datasets by the topics that are common in the content. With the email dataset, we compared both algorithms like LDA and DMM, which were used frequently before, and the algorithms that enable topic modeling from short texts that have gained popularity in recent years.

After the emails are clustered into a suitable number of topics and the keywords of the topics are determined, the results are transferred to the admin template generation panel and made ready for template creation, which can be seen in Figure 6. This system suggests the templates, which are edited and shared by the admin, can be used by detecting the topic of a users' new emails while it is being written. The user can use the template depending on his / her request.

## REFERENCES

[1] Yin, J., & Wang, J. (2014). A dirichlet multinomial mixture model-based approach for short text clustering. *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '14.* doi:10.1145/2623330.2623715

[2] Li, C., Wang, H., Zhang, Z., Sun, A., & Ma, Z. (2016). Topic Modeling for Short Texts with Auxiliary Word Embeddings. *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '16.* doi:10.1145/2911451.2911499

[3] Li, C., Duan, Y., Wang, H., Zhang, Z., Sun, A., & Ma, Z. (2017). Enhancing Topic Modeling for Short Texts with Auxiliary Word Embeddings. *ACM Transactions on Information Systems, 36*(2), 1-30. doi:10.1145/3091108

[4] Nguyen, D. Q., Billingsley, R., Du, L., & Johnson, M. (2015). Improving Topic Models with Latent Feature Word Representations. *Transactions of the Association for Computational Linguistics, 3*, 299-313. doi:10.1162/tacl_a_00140

[5] Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3, 993–1022. doi: http://dx.doi.org/10.1162/jmlr.2003.3.4-5.993

[6] Yan, X., Guo, J., Lan, Y., & Cheng, X. (2013). A biterm topic model for short texts. *Proceedings of the 22nd International Conference on World Wide Web - WWW '13.* doi:10.1145/2488388.2488514

[7] Zuo, Y., Zhao, J., & Xu, K. (2015). Word network topic model: A simple but general solution for short and imbalanced texts. *Knowledge and Information Systems, 48*(2), 379-398. doi:10.1007/s10115-015-0882-z

[8] Zuo, Y., Wu, J., Zhang, H., Lin, H., Wang, F., Xu, K., & Xiong, H. (2016). Topic Modeling of Short Texts: A Pseudo-Document View. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. doi:10.1145/2939672.2939880

[9] Quan, X., Kit, C., Ge, Y., & Pan, S. (2015). Short and Sparse Text Topic Modeling via Self-Aggregation. In *Proceedings of the 24th International Conference on Artificial Intelligence* (pp. 2270–2276). AAAI Press.

[10] M. Dredze, H. M. Wallach, D. Puller, and F. Pereira, "Generating summary keywords for emails using topics," *Proceedings of the 13th international conference on Intelligent user interfaces - IUI 08*, 2008.

[11] T. Ayodele and S. Zhou, "Applying Machine learning Algorithms for Email Management," *2008 Third International Conference on Pervasive Computing and Applications*, 2008.

[12] T. Ayodele, S. Zhou, and R. Khusainov, "Evolving email clustering method for email grouping: A machine learning approach," *2009 Second International Conference on the Applications of Digital Information and Web Technologies*, 2009.

[13] S. Joty, G. Carenini, G. Murray, Ng. Raymond. (2010). Exploiting Conversation Structure in Unsupervised Topic Segmentation for Emails.. 388-398.

[14] H. Yang, J. Luo, M. Yin, and Y. Liu, "Automatically Detecting Personal Topics by Clustering Emails," *2010 Second International Workshop on Education Technology and Computer Science*, 2010.

[15] S. Pan, M. X. Zhou, Y. Song, W. Qian, F. Wang, and S. Liu, "Optimizing temporal topic segmentation for intelligent text visualization," *Proceedings of the 2013 international conference on Intelligent user interfaces - IUI 13*, 2013.

[16] H. Hong and T.-S. Moh, "Effective topic modeling for email," *2015 International Conference on High-Performance Computing & Simulation (HPCS)*, 2015.

[17] Qiang, J., Li, Y., Yuan, Y., Liu, W., & Wu, X. (2018). STTM: A Tool for Short Text Topic Modeling. ArXiv, abs/1808.02215.

[18] McCallum, Andrew Kachites. "MALLET: A Machine Learning for Language Toolkit." http://mallet.cs.umass.edu. 2002.

[19] W. Zhang, A. Ahmed, J. Yang, V. Josifovski, and A. J. Smola, "Annotating Needles in the Haystack without Looking," *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 15*, 2015.

[20] J. Proskurnia, M.-A. Cartright, L. Garcia-Pueyo, I. Krka, J. B. Wendt, T. Kaufmann, and B. Miklos, "Template Induction over Unstructured Email Corpora," *Proceedings of the 26th International Conference on World Wide Web*, Mar. 2017.

[21] M. Whittaker, N. Edmonds, S. Tata, J. B. Wendt, and M. Najork, "Online template induction for machine-generated emails," *Proceedings of the VLDB Endowment*, vol. 12, no. 11, pp. 1235–1248, Jan. 2019.

[22] Mikolov, T., Chen, K., Corrado, G.S., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. CoRR, abs/1301.3781.

[23] Code.google.com. 2020. *Google Code Archive - Long-Term Storage For Google Code Project Hosting.* [online] Available at: https://code.google.com/archive/p/word2vec/ [Accessed 15 July 2020].

[24] Wikimedia. (n.d.) 2020, [online] Available at: https://dumps.wikimedia.org/other/wikibase/wikidatawiki/ [Accessed 15 July 2020].

GCRIS