Chaipornkaew, Piyanuch; Prexawanprasut, Takorn; Chang, Chia-Lin; McAleer, Michael

Working Paper
A Generalized Email Classification System for Workflow Analysis

Tinbergen Institute Discussion Paper, No. 17-066/III

Provided in Cooperation with:
Tinbergen Institute, Amsterdam and Rotterdam

Suggested Citation: Chaipornkaew, Piyanuch; Prexawanprasut, Takorn; Chang, Chia-Lin; McAleer, Michael (2017) : A Generalized Email Classification System for Workflow Analysis, Tinbergen Institute Discussion Paper, No. 17-066/III, Tinbergen Institute, Amsterdam and Rotterdam

This Version is available at:
http://hdl.handle.net/10419/177634

Standard-Nutzungsbedingungen:
Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebräuch gespeichert und kopiert werden.
Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.
Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:
Documents in EconStor may be saved and copied for your personal and scholarly purposes.
You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.
If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.
A Generalized Email Classification System for Workflow Analysis

Piyanuch Chaipornkaew¹
Takorn Prexawanprasut²
Chia-Lin Chang³
Michael McAleer⁴

1: College of Innovative Technology and Engineering, Dhurakij Pundit University, Thailand
2: College of Innovative Technology and Engineering, Dhurakij Pundit University, Thailand
3: Department of Applied Economics and Department of Finance, National Chung Hsing University, Taiwan
4: Department of Quantitative Finance, National Tsing Hua University, Taiwan; Discipline of Business Analytics, University of Sydney Business School, Australia; Econometric Institute, Erasmus School of Economics, Erasmus University Rotterdam, The Netherlands; Department of Quantitative Economics, Complutense University of Madrid, Spain; Institute of Advanced Sciences, Yokohama National University, Japan
Tinbergen Institute is the graduate school and research institute in economics of Erasmus University Rotterdam, the University of Amsterdam and VU University Amsterdam.

Contact: discussionpapers@tinbergen.nl

More TI discussion papers can be downloaded at the Tinbergen Site

Tinbergen Institute has two locations:

Tinbergen Institute Amsterdam
Gustav Mahlerplein 117
1082 MS Amsterdam
The Netherlands
Tel.: +31(0)20 598 4580

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31(0)10 408 8900
A Generalized Email Classification System for Workflow Analysis*

Piyanuch Chaipornkaew  
College of Innovative Technology and Engineering  
Dhurakij Pundit University  
Bangkok, Thailand  

Takorn Prexawanprasut  
College of Innovative Technology and Engineering  
Dhurakij Pundit University  
Bangkok, Thailand  

Chia-Lin Chang  
Department of Applied Economics  
Department of Finance  
National Chung Hsing University  
Taichung, Taiwan  

Michael McAleer  
Department of Quantitative Finance  
National Tsing Hua University, Taiwan  
and  
Discipline of Business Analytics  
University of Sydney Business School, Australia  
and  
Econometric Institute, Erasmus School of Economics  
Erasmus University Rotterdam, The Netherlands  
and  
Department of Quantitative Economics  
Complutense University of Madrid, Spain  
and  
Institute of Advanced Sciences  
Yokohama National University, Japan  

Revised: July 2017

* The authors would like to thank the Executive Vice-President of Finish International Freight Co.Ltd, as well as two anonymous companies which cannot be mentioned because of confidentiality. The companies provided the useful information to conduct this research. Thanks also to Khun Natthicha Phonjan and Khun Sariporn Plipon, who assisted in manually classifying the emails. It is appreciated that the business data provided by the three selected businesses are sensitive, and will not be disclosed or used for any purpose other than the research for the paper. Corresponding author: Takorn Prexawanprasut (takorn.pre@dpu.ac.th)
Abstract

One of the most powerful internet communication channels is email. As employees and their clients communicate primarily via email, much crucial business data is conveyed via email content. Where businesses are understandably concerned, they need a sophisticated workflow management system to manage their transactions. A workflow management system should also be able to classify any incoming emails into suitable categories. Previous research has implemented a system to categorize emails based on the words found in email messages. Two parameters affected the accuracy of the program, namely the number of words in a database compared with sample emails, and an acceptable percentage for classifying emails. As the volume of email has become larger and more sophisticated, this research classifies email messages into a larger number of categories and changes a parameter that affects the accuracy of the program. The first parameter, namely the number of words in a database compared with sample emails, remains unchanged, while the second parameter is changed from an acceptable percentage to the number of matching words. The empirical results suggest that the number of words in a database compared with sample emails is 11, and the number of matching words to categorize emails is 7. When these settings are applied to categorize 12,465 emails, the accuracy of this experiment is approximately 65.3%. The optimal number of words that yields high accuracy levels lies between 11 and 13, while the number of matching words lies between 6 and 8.

Keywords: Email, business data, workflow management system, business transactions.

JEL: J24, O31, O32, O33.
1. Introduction

Information and communication technology has been developed significantly in recent years. The technology eliminates the wall of distance and connects people more closely than ever. The technology also supports many businesses to gain competitive advantages. Owing to this technology, large numbers of organizations are able to operate their business at lower costs and with a higher competitive advantage. As a result, many organizations attempt to acquire this on-time and accurate information. One of the most powerful tools in business is email, which is a fundamental and indispensable communication channel for every organization in the modern age.

In recent decades, the number of startup companies has increased dramatically. Two of the authors have participated in three start-up companies related to the import/export sector. These new start-ups established their own businesses by separating themselves from their former companies. After the initial study, it was found that startup companies needed to manage a large number of daily documents/emails because startup businesses contacted their customers and employees primarily via email. The employees also used these emails, which were stored in the mail server, as a database. For example, when employees wanted to find specific data, emails were the first place for seeking information.

In the first stage of starting their businesses, the number of emails was not large. However, when the scale of business expanded, the number of emails increased. The business owners needed applications to manage their company activities, a problem that could be solved primarily by software applications, such as the workflow management system. However, the cost of this software is rather high, and may not be appropriate for startup companies, so that alternative approaches to solve the problem were needed.

For the initial investigation, 12,465 of emails were selected from the three startup companies because they were written in English. As the employees in the selected companies wrote emails in two languages, namely English and Thai, only emails that were written in English were taken into consideration as the sentences in English are easier to separate into words than corresponding emails in Thai. By investigating some of these emails, some keywords specified the type of work, such as
sales, transportation, billing, or shipping, which can be used as initial guidelines to conduct the classification models.

The purpose of this paper is to define the categories of email and extract business data for a workflow management system.

The remainder of the paper is as follows. Section 2 provides a literature review, Section 3 describes the materials and methods, Section 4 presents the data analysis, Section 5 illustrates the results and discussion, and Section 6 provides some concluding comments.

2. Literature Review

There is much research that mentions the clustering and classification of email content, and many objectives to conduct research for email classification problem, such as: distinguishing between personal and machine-generated email [1]; classifying emails for contact centers [2]; classifying emails for automated service handling [3]; and classifying emails for social network analysis [4]. As regards classification techniques, there are also many methods applied to email classification, such as mining-based approaches [5], supervised learning algorithms [6], co-training technique [7,8], co-training with a Single Natural Feature Set [9], and regression-based approaches [10].

One of the interesting topics is by Alsmadia and Alhamib [11]. The authors illustrate that the best algorithm to perform email clustering and classification is Ngram. Their sets of emails were in the form of a large text collection, which fits with the Ngram algorithm, and the algorithm best fits the bi-language text. They conducted an experiment based on emails in both English and Arabic. The major challenge of their future work was that email servers or applications should include different types of pre-defined folders. The general pre-defined folders could be mailbox, sent, or trash, among others. Moreover, email servers or applications could allow users to add new folders for specific purposes, based on their Ngram algorithm.

Further research on email classification is by Katakis, Tsoumakas, and Vlahavas [12]. They state that Machine Learning and Data Mining could be used as tools to automate email managing tasks, which could be far superior to other conventional solutions. They discuss the particularity of email content, and what special treatment it requires. In addition, there are some interesting email mining
applications, like mail categorization, summarization, automatic answering, and spam filtering. In their experiments, they created an application to classify email based on several techniques, such as the Naïve Bayes Classifier and Support Vector Machines.

Ayodele, Khusainov, and Ndzi [13] present the design and implementation of a system to group, and summarize email messages. Their system considers the subject and content of email messages to classify emails based on user activities, and produces summaries of each incoming message with an unsupervised learning approach. They claim that their framework could solve the problems of email overload, congestion, difficulties in prioritizing, and difficulties in finding previously archived messages in the email server.

Another interesting topic is email grouping and summarization. Ayodele, Zhou, and Khusainov [14] present the design and implementation of an application to categorize and summarize email content. Their system extracts the subject and content of email messages for classification based on user activities to auto-generate a summary of each incoming message. They state that their framework could solve problems such as email overload, difficulties in prioritizing, and email congestion. Their framework also performs successful processing of new incoming messages.

Another interesting concept is automated email activity management, as in Kushmerick and Lau [15], who develop email applications that provide high-level support for structured activities in e-commerce. They define formal activities as finite-state automata, which correspond to the status of the process, and where transitions represent messages sent between participants. They propose several unsupervised machine learning algorithms, and evaluate a collection of e-commerce emails.

Schuff, Turetken, D'Arcy, and Croson [16] also discuss email classification. They implement effective e-mail management tools, which treat messages as useful information. This tool could economize on scarce cognitive resources at the expense of relatively cheap additional CPU power, disk capacity, and network bandwidth. In addition, they claim that their application provides automatic filtering, clustering, and a new user interface. Their system employs a large number of emails as an effective knowledge management tool, rather than as a source of information overload.

Email classification is discussed in Prexawantprasut and Chaipornkaew [17]. The research classifies email into four categories, namely sales, shipping, billing, and transportation. Two parameters are
applied for the classification system, namely the number of words in a database compared with the sample emails, and an acceptable percentage to classify emails. The accuracy of classification is determined to be approximately 73.6%.

Chaipornkaew, Prexawanprasut, and McAleer [18] discuss email extraction for workflow management system. In order to extract data, there are four criteria which are applied. Fifteen cases of alternative criteria to extract data are analyzed. The results show that when criteria numbers 2 and 4 are considered, email extraction accuracy is at the highest level. However, when the highest accuracy level occurs, the number of blanks fields is also high. According to user requirements, the number of blank fields should be at a low level. Therefore, the paper suggests that all four criteria should be considered to provide both an acceptable percentage of blank fields and also accuracy level.

3. Materials and Methods

The paper is planned in two phases, as shown in Figure 1. First, 1260 emails are selected randomly from the server to be used as training data for the system. These emails are then classified manually by employees into seven categories, namely (1) Sales, (2) Agent, (3) Shipping, (4) Customs, (5) Billing, (6) Packing and Moving, and (7) Insurance. The sentences in emails are separated into words, which are counted, as shown in Figure 2. These results are stored in the database, which is applied for email classification rules.

In order to test the defining rules, a further 12,465 emails are selected from the server. When these rules are accepted, the rest of the emails in the server are processed by the program. After the classification is processed, all emails are assigned to suitable categories, and then all the data are prepared for the second phase of the email classification system.

The second phase is to extract the classified emails, which are processed from the first phase. As in investigating the selected emails, there are key characteristics which can be represented as relationships. For example, the document number could be a key characteristic to define the relationships among the email messages. The program first reorders emails based on time in each
category, then extracts data based on their characteristics. The final stage is to create a workflow management system from the extracted data.

4. Data Analysis

The first stage is to export all emails from the email server and format them in a text file, which is then imported to the program. The program first separates words in a text file. As the selected emails are in English, the algorithm to separate the words is the use of spaces. The words from the separation process are counted and stored in a database. The database stores all results which are all words, and their frequencies as shown in Table 1.

The research classifies 12,465 emails into seven categories, namely (1) Sales, (2) Agent, (3) Shipping, (4) Customs, (5) Billing, (6) Packing and Moving, and (7) Insurance. The mechanism is implemented based on the words found in emails compared with the words in the database for each category. Two parameters are considered in this experiment. The first parameter is the number of words in the database. For example, in order to gain greater accuracy in the classification, we need to determine whether the first 3 or 5 words in the database should be considered. The second parameter is the number of matching words that provides the highest accuracy to determine the category of email.

According to the data in Table 2, some email could not be classified because the number of matching words is less than the specified criteria. In this case, the second criterion is the first 5 words in a database. In order to obtain better results, these two criteria may need to be refined. As shown in Table 3, the first 10 words in a database are considered instead of the first 5 words.

The number of matching words is set at 5 in Table 2, and set at 4 in Table 3. As a result, only two groups of output in Tables 2 and 3 are the same. The first difference is the No. 2 group of emails. In Table 2, Email No. 2 could be either Sales or Packing and Moving, but it is concluded to be Packing and Moving group in Table 3. The second difference is the No. 3 group of emails, which could not be grouped in Table 2, but could be defined as Agent in Table 3. The third difference is the No. 5 group of emails, which is defined as Insurance group in Table 2, while in Table 3 it is concluded to be Shipping.
The empirical data from both Tables 2 and 3 demonstrate that are two main factors that affect the grouping results. The first factor is the number of words in the database to be considered, while the second factor is the number of matching words. Therefore, another 12,465 emails are collected to test the program by changing the criteria for these two factors, with the empirical results shown in Figure 3.

5. Results and Discussion

The results shown in Figure 3 illustrate that the accuracy levels change when the number of words in the database and the number of matching words change. The purpose of the paper is to discover suitable parameter values, namely: (1) the number of words in the database to be considered; and (2) the number of matching words. The number of words in a database to be considered is adjusted from 5 to 20, while the number of matching words are adjusted from 1 to 20.

According to the results in Figure 3, the highest accuracy level of email classification occurs when the number of words in a database is 11 and the number of matching words is 7. Therefore, these criteria are applied in the program. The program then classified the other 12,465 emails into seven groups, namely: (1) Sales, (2) Agent, (3) Shipping, (4) Customs, (5) Billing, (6) Packing and Moving, and (7) Insurance, as shown in Table 4.

According to Table 4, the program could not categorize all the emails because some emails do not meet the acceptable criteria. The program is able to define only 9,972 emails from a total of 12,465 emails, which represents 80% of the total. There are 2,493 emails which could not be categorized in the experiment. In order to improve the program efficiency, other factors could be concerned. One possible factor could be the importance level of each word (the weight of each word) in a database. For example, words that are found most frequently in emails should be placed at a higher level of importance than those that are found less frequently.

When the first phase is completed, all emails are already classified into groups (Sales, Agent, Shipping, Customs, Billing, Packing and Moving, and Insurance). The next phase is to analyze the characteristics of the emails. Key characteristics are defined by employees. The program collects these characteristics, which are applied for data extraction. The program reorders the events based on time in each category, as shown in Figure 4.
The last stage is to extract the specified data based on their characteristics. As the characteristics of data are in many forms, the extracted data can vary substantially. One example of data which are extracted based on Document Number (FWO0018) is shown in Figure 5. According to the results, all the details concerned with Document Number (FWO0018) are well summarized. The data that are extracted will be stored in a database, which will be implemented for a workflow management system.

6. Conclusion

According to the experiments, the accuracy level of email classification depends on two factors, namely the number of words to be considered in a database, and the number of matching words. After testing the program with different values for these two factors, the results show that the optimal value for the number of words in a database is 11, while the number of matching words is 7. The results also illustrate that high accuracy levels fall in the range of the number of words lying between 11 and 13, while the range of the number of matching words lies between 6 and 8.

As mentioned earlier, the experiments select all emails in English, so some words need to be neglected. Examples of words which should not be considered are ‘and’, ‘not’, ‘thanks’, ‘regards’, and ‘please’. As these words could be found in most emails, they should not be included in the program. As these words could not be used as criteria to classify email, a more sophisticated program should be developed to ignore these words before processing the email classifications.

In investigating email content, there are specific words that should not be used as criteria in email classification. Examples of these words are FREIGHTLINKS, STARSHIP, and HERMESINTL. As these words are actual customer names, they should be defined as customer names in the database, and are excluded from the criteria for email classification in the first phase. However, these specific data are the key characteristics for the second phase of the research. The data with their characteristics are applied to extract data, which are used for the workflow management system.

The generalized email classification system for workflow analysis has been shown to work well in the experiments, with a high degree of accuracy.
Figure 1

Two Phases of the Email Classification System
Figure 2

Example of Results from the Word Separation Process
<table>
<thead>
<tr>
<th>Number of Words</th>
<th>Number of Matching Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20</td>
<td></td>
</tr>
<tr>
<td>5 13.7 27.9 30.6 30.1 29.9</td>
<td>20 24 32.1 32.5 33.1</td>
</tr>
<tr>
<td>6 18.2 19.5 20.1 23.6 36.8 32.3 23.3</td>
<td>8 16.2 17.5 18.5 32.5 40.1 23.2 20.1 18.3</td>
</tr>
<tr>
<td>9 15.2 23.6 38.8 42.6 42.1 39.6 35.3 35.1 23.5</td>
<td>10 19.2 19.2 35.5 48.5 47.8 50.1 38.5 26.5 28.2 20.5</td>
</tr>
<tr>
<td>11 14.6 25.5 28 36.3 45.1 56.8 65.5 52.2 35.1 18.2 10.8</td>
<td>12 15.2 15.5 20.4 25.7 40.6 58 58.1 60.1 39 20.3 10.3 12.8</td>
</tr>
<tr>
<td>13 15.7 19.6 21.5 26 47.5 63.5 54.9 62.1 34 25.1 19.5 15.1 15.6</td>
<td>14 14.3 20.1 12.1 15.6 20.7 32.2 34 45.6 28.9 29.4 18.8 12.3 10.2 8.75</td>
</tr>
<tr>
<td>15 14.5 15.7 13.3 12.1 15.6 18.3 20.3 14.5 13.5 12.5 10.8 8.65 8.5 7.85 12.3</td>
<td>16 12 19 15.7 13.2 12.6 11.9 10.8 18.5 15.2 12.3 14.6 13.8 17.4 12.1 11.5 10.8</td>
</tr>
<tr>
<td>17 6.02 19.5 12.3 11.9 10.6 8.72 14.3 13.2 12.6 11.6 13.2 17.6 15.5 13.2 11.2 9.25 10.8</td>
<td>18 10.6 12 10.2 8.92 9.95 8.75 12.2 15.3 10.8 18.5 12.2 12 8.16 7.13 10.6 7.75 10.2 9.8</td>
</tr>
<tr>
<td>19 8.12 9.87 10.2 13.9 10.3 13.8 15.5 18 15.9 13.5 13 13.3 14.1 8.54 12.3 8.63 8.72 9.58 9.98</td>
<td>20 9.5 7.14 12.5 8.12 10.2 11.5 15.5 13.9 12.5 11.6 10.1 13.5 12.2 10.5 8.56 6.23 8.59 6.54 7.63 9.72</td>
</tr>
</tbody>
</table>

**Figure 3**

Accuracy (%) of Email Classification
Figure 4

Program Results after Grouping, Event Ordering, and Inclusion of Email Characteristics
Figure 5

Example of Extracted Data from Phase 2, Based on Document Number

<table>
<thead>
<tr>
<th>Related issues from mail number 4281,4292,4297,4320....</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Document Number 1</strong>: FWO0018-00008</td>
</tr>
<tr>
<td><strong>Document Number 2</strong>: FWO0018-00017</td>
</tr>
<tr>
<td><strong>Document Number 3</strong>: FWO0018-00019</td>
</tr>
</tbody>
</table>

**Icon Global Logistics / HO20-1600078**

**Shipper**: CHIEF

**Cconsignee**: POP MEDIC

**PO No**: MKC 01-2017

**Volume**: 46 cartons / 549.8 kg / 4.

**Vessel/vooy**: NORDOCELOT V-201Q

**Consignee**: 3days

**Closing date**: 17/07/14

**ETD**: 17/07/18

**ETA**: 17/07/25

**Remarks**: Subject to local charge both of sides

*add related documents*

*add related participants*

*add related events*

*add related remarks*
### Table 1

**Top 15 Words in Emails in 7 Categories**

<table>
<thead>
<tr>
<th>Sales</th>
<th>Agent</th>
<th>Shipping</th>
<th>Customs</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Word</strong></td>
<td><strong>Frequency</strong></td>
<td><strong>Word</strong></td>
<td><strong>Frequency</strong></td>
</tr>
<tr>
<td>agent</td>
<td>112</td>
<td>#NAME of CUS</td>
<td>188</td>
</tr>
<tr>
<td>volume</td>
<td>91</td>
<td>arrange</td>
<td>165</td>
</tr>
<tr>
<td>#NAME of CUS</td>
<td>88</td>
<td>ETA</td>
<td>150</td>
</tr>
<tr>
<td>product</td>
<td>72</td>
<td>delivery</td>
<td>112</td>
</tr>
<tr>
<td>shipment</td>
<td>60</td>
<td>#NAME of CITY</td>
<td>94</td>
</tr>
<tr>
<td>#NAME of CITY</td>
<td>58</td>
<td>import</td>
<td>86</td>
</tr>
<tr>
<td>process</td>
<td>55</td>
<td>items</td>
<td>81</td>
</tr>
<tr>
<td>confirm</td>
<td>52</td>
<td>#NAME of PORT</td>
<td>75</td>
</tr>
<tr>
<td>week</td>
<td>48</td>
<td>warehouse</td>
<td>53</td>
</tr>
<tr>
<td>#Date format</td>
<td>31</td>
<td>service</td>
<td>50</td>
</tr>
<tr>
<td>D/O</td>
<td>28</td>
<td>update</td>
<td>48</td>
</tr>
<tr>
<td>packing list</td>
<td>25</td>
<td>port</td>
<td>39</td>
</tr>
<tr>
<td>#NAME of PORT</td>
<td>22</td>
<td>shipping</td>
<td>31</td>
</tr>
<tr>
<td>attach</td>
<td>18</td>
<td>scheduled</td>
<td>21</td>
</tr>
<tr>
<td>request</td>
<td>16</td>
<td>#Date format</td>
<td>12</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Billing</th>
<th>Packing and Moving</th>
<th>Insurance</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Word</strong></td>
<td><strong>Frequency</strong></td>
<td><strong>Word</strong></td>
</tr>
<tr>
<td>consignee</td>
<td>125</td>
<td>loading</td>
</tr>
<tr>
<td>shipper</td>
<td>111</td>
<td>destination</td>
</tr>
<tr>
<td>document</td>
<td>94</td>
<td>package</td>
</tr>
<tr>
<td>revise</td>
<td>89</td>
<td>carrier</td>
</tr>
<tr>
<td>#NAME of CUS</td>
<td>84</td>
<td>loader</td>
</tr>
<tr>
<td>scheduled</td>
<td>74</td>
<td>#Date format</td>
</tr>
<tr>
<td>departed</td>
<td>62</td>
<td>co-loader</td>
</tr>
<tr>
<td>service</td>
<td>50</td>
<td>departed</td>
</tr>
<tr>
<td>#Date format</td>
<td>48</td>
<td>ETD</td>
</tr>
<tr>
<td>arrived</td>
<td>42</td>
<td>arrived</td>
</tr>
<tr>
<td>#NAME of PORT</td>
<td>38</td>
<td>scheduled</td>
</tr>
<tr>
<td>shipment</td>
<td>31</td>
<td>#NAME of PORT</td>
</tr>
<tr>
<td>notice</td>
<td>25</td>
<td>shipment</td>
</tr>
<tr>
<td>booking</td>
<td>22</td>
<td>worker</td>
</tr>
<tr>
<td>approval</td>
<td>18</td>
<td>condition</td>
</tr>
</tbody>
</table>
### Table 2
Grouping Results Based on Top 5 Words and 5 Acceptable Number of Matching Words

<table>
<thead>
<tr>
<th>No. of Emails</th>
<th>Number of Matching Words</th>
<th>Grouping result</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sales</td>
<td>Agent</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

**Note:** In the case of email no. 2, it falls into either Sales or Packing and Moving category. The research could not conclude whether it should be in the Sales or Packing and Moving group. This issue should be clarified in future research.
### Table 3
Grouping Results Based on Top 10 Words and 4 Acceptable Number of Matching Words

<table>
<thead>
<tr>
<th>No. of Emails</th>
<th>Sales</th>
<th>Agent</th>
<th>Shipping</th>
<th>Customs</th>
<th>Billing</th>
<th>Packing and Moving</th>
<th>Insurance</th>
<th>Grouping result</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>Sales</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>Packing and Moving</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>Agent</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>4</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>Billing</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>4</td>
<td>8</td>
<td>0</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>Shipping</td>
</tr>
</tbody>
</table>
Table 4
Number of Emails in Each Category

<table>
<thead>
<tr>
<th>Sales</th>
<th>Agent</th>
<th>Shipping</th>
<th>Customs</th>
<th>Billing</th>
<th>Packing and Moving</th>
<th>Insurance</th>
<th>Unclassified</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,994</td>
<td>1,623</td>
<td>1,246</td>
<td>1,121</td>
<td>1,371</td>
<td>1,745</td>
<td>872</td>
<td>2,493</td>
<td>12,465</td>
</tr>
</tbody>
</table>
References


