data

MDPI

Data Descriptor

eMailMe: A Method to Build Datasets of Corporate Emails in Portuguese

Akira A. de Moura Galvão Uematsu * and Anarosa A. F. Brandão

Engenharia de Computação e Sistemas Digitais, Escola Politécnica-Universidade de São Paulo, Av. Prof. Luciano Gualberto, São Paulo 05508-010, Brazil

* Correspondence: akira.uematsu@usp.br; Tel.: +55-11-3206-1613

Abstract: One of the areas in which knowledge management has application is in companies that are concerned with maintaining and disseminating their practices among their members. However, studies involving these two domains may end up suffering from the issue of data confidentiality. Furthermore, it is difficult to find data regarding organizations processes and associated knowledge. Therefore, this paper presents a method to support the generation of a labeled dataset composed of texts that simulate corporate emails containing sensitive information regarding disclosure, written in Portuguese. The method begins with the definition of the dataset's size and content distribution; the structure of its emails' texts; and the guidelines for specialists to build the emails' texts. It aims to create datasets that can be used in the validation of a tacit knowledge extraction process considering the 5W1H approach for the resulting base. The method was applied to create a dataset with content related to several domains, such as Federal Court and Registry Office and Marketing, giving it diversity and realism, while simulating real-world situations in the specialists' professional life. The dataset generated is available in an open-access repository so that it can be downloaded and, eventually, expanded.

Keywords: knowledge management; knowledge acquisition; tacit knowledge; 5W1H; natural language processing

Citation: Uematsu, A.A.d.M.G.; Brandão, A.A.F. eMailMe: A Method to Build Datasets of Corporate Emails in Portuguese. *Data* **2023**, *8*, 127. https://doi.org/10.3390/data8080127

Academic Editors: Francesco M. Donini and Joaquín Torres-Sospedra

Received: 7 May 2023
Revised: 18 July 2023
Accepted: 26 July 2023
Published: 31 July 2023

1. Introduction

Knowledge management (KM) is concerned with the representation, organization, acquisition, creation, usage, and evolution of knowledge in its many forms [1]. In large corporations, it is of strategic importance that all its members have knowledge about the processes that are performed there, about where they can find certain types of information, directions that are being taken, etc. Therefore, it is expected that many companies have areas dedicated to the application of KM in their various spheres. Moreover, for these sectors to keep evolving, it is natural that research be carried out in this area. However, given the nature of data involved in these studies, many of them cannot be disclosed due to confidentiality issues.

In this article, we present a method to support the development of labeled datasets of corporate emails written in Brazilian Portuguese. This dataset was designed for tacit knowledge extraction. Tacit knowledge presents some characteristics that make it different from others, such as being personal; being difficult to articulate and contextualize; being specific to the area; and often being more easily shared through conversations or narratives. In short, it is knowledge that the employee has but has difficulties in expressing [2]. And this knowledge can be important if we think in terms of specialists, that is, those professionals who have remained satisfactorily in their activity for a long period. This person's tacit knowledge has professional experiences, which could make the efficiency of activities improve and which are not described anywhere. It would not be in anyone's interest for this type of knowledge to be lost with her when she eventually leaves the company.

Knowledge can be expressed by using the 5W1H standard, which works as a label for the elements of a sentence. This acronym means "Who" did "What", "When", "Where", "Why", and "How"; therefore, we are left with a pair of keyword and value that directly summarizes essential elements for understanding a given text. In the example *"John went to the museum yesterday"*, we can identify four of the W's; John is "Who", went is "What", museum is "Where", and yesterday is "When". Thus, it is natural to imagine a direct application of this standard in the field of journalism, as we can see in [3] with the news summary. But also, it has applications in many areas, like in KM itself. Interviews with specialists in a given professional area represent a great opportunity to extract and store knowledge. According to Supnitchaisiri, Natakuatoong, and Sinthupinyo [4], one of the ways to carry out this questioning is through 5W1H. The identification of these elements can contribute to the retention process of tacit knowledge from specialists. Another area where this standard is applied is criminal investigations where, unlike our case, the information comes from multiple documents and an auxiliary module is used to identify technical criminal terms that are contained in the input data [5].

Based on this standard, the method guides the construction of email text content following grammatical rules that clarify, for example, when an action occurred or why something happened. In addition, such content is created by knowledge domain specialists in order to mimic the behavior of professional communication inside corporations.

Corporate email datasets like this could be obtained by direct server access, sender identification, and sending date labeling. Nevertheless, it is our claim that providing a method to support the development of a dataset whose content simulates email texts like the aforementioned is of interest for the KM research community. In fact, the confidential nature of the content that may exist in this type of text, not only related to the company but also regarding who wrote it, would prevent access to them and research on the theme. Therefore, we understand that one of our main contributions is the availability of the dataset resulting from the methodology presented in this work since, to the best of our knowledge, datasets that gather texts of a personal nature written in Portuguese are not openly available. As we see in [6], despite Portuguese being a language with millions of native speakers, there are not many datasets available to work with in some areas, much less when the texts deal with professional content. In our case, this professional content was ensured through the contributions of people with different experiences and professional activities in content creation.

Related Work

Having access to labeled datasets may not be a simple task. As we can see in [7], where a spam identification system is created for a specific language and it is concluded that this type of labeling depends on human intelligence, due to meaning interpretations of the same word within the context. In our article, data labeling, is carried out by experts in the field using the methodology described, as these are emails with personal content and therefore cannot be reproduced using computational processes.

Consulting specialists from a professional area, such as those recruited to follow our methodology, is a practice that was also observed in [8], where interviews with experts in data creation in the field of Machine Learning were conducted and the difficulties in obtaining quality data were identified. One of the conclusions of the article is that, first it's necessary to identify and record the ways in which people generate data, and their efforts to ensure that these data are of good quality. However, in our case, the entire creation process is already registered through the methodology.

Existing techniques in the literature can be used to build datasets. In [9] we have the description of the construction of a labeled images dataset from roofs, according to the CityGML standard, considering the diversity of roofs found in the city of Sofia, Bulgaria. However, in our case we describe a new method for constructing a labeled dataset of corporate emails with respect to whether or not they contain implicit knowledge.

Finally, the use of simulated data sets ends up being an unavoidable practice in some areas, due to the restrictions of the contexts where they are inserted. As we see in [10], where a dataset was generated through a daily activities simulation tool, in a virtual house equipped with a smart home environment. Unlike our case where domain experts were asked to create email texts simulating content that would be exchanged in corporate environments. Then, as previously mentioned, obtaining some datasets directly can end up being difficult. As in this case of smart homes, regarding the time to capture data and the cost of installing sensors, or as in our case, where there is a question of confidentiality of the content that would exist in a personal email.

This paper is structured in five sections. Section 2 detail the specific characteristics that the dataset presents and section 3 presents the guidelines that must be followed for constructing the email content, besides, provide more detail about the eMailMe application considering three specialists of diverse domains. Section 4 details the best way to handle with the dataset and, finally section 5 presents some discussion and conclusions.

2. Data Description

The dataset has some specific characteristics that will be presented in this section and were defined through the creation project. This includes content management as well as the consolidation of texts in a format suitable for processing.

As mentioned in the introduction, the dataset content must simulate texts from corporate emails produced by specialists in a given area. In addition, the labeling, which will be further detailed in Section 3.4.3, would indicate which texts embed tacit knowledge within them, the sending date, and identifications for the email groups and for each one of the emails texts.

2.1. File Format

The dataset must bring some complementary information to label the data. In order to create a sort of database in a single file without any database management system, we chose to build the dataset in a spreadsheet format. This means that the file extension could be any of the following extensions: XLS, XLSX, XLSM, XLSB, ODF, or ODS. Also, as the texts will be written in Brazilian Portuguese, we define the character set in UTF-8 to contain the diacritical marks used in that language. There are other character encodings that also contain these marks; however, our choice was based on the one with the most widespread use. And as long as the encoding used is declared when working with the file, no problem should occur.

2.2. Spreadsheet Contents in the File

Each spreadsheet must follow a preset layout, described at Table 1, and is expected to contain texts related to the same context.

As can be seen in Table 1, the spreadsheet contains a header of three lines. The first line contains a base identification number (`base_id`), and in the second line we specified the context (`area`), providing information on which class of corporation this type of content would be inserted in; for example, we could create texts about a specialist who works in a bank, hospital, law firm, etc. In the third line, it is specified whenever the text contains (or not) simulated tacit knowledge (`knowledge`). This line represents the label that can be used for tacit knowledge extraction and acquisition.

After the header, we can find the dataset entries distributed in three columns. The first column contains the e-mail identification number; as seen above, we started from id 1 and increase by one with each new email. The second column contains the date on which the e-mail would have been sent, according to format YYYY-MM-DD; above, we see the days in a sequence within the month, but we will not always have a line for every day, as this is not mandatory. Finally, the third column contains the text of the e-mail itself; in the table are some examples with the content "E-mail's 1 text" as a reminder that texts are written in Portuguese.

Table 1. Created spreadsheet layout.

base_id	1	
area	Name of the area in which the texts are inserted	
knowledge	True	
1	2022-05-09	Texto do e-mail 1
⋮	⋮	⋮
N	2022-05-09	Texto do e-mail N

When inserting elements related to "When" in the sentence, it must be observed if the informed date in the first column of the file is coherent. For example, the phrase *"Yesterday, 9 May 2023, payroll was processed"* would not make sense if, in the email sent date column, we enter 2023-05-09. After all, it is necessary to reproduce what we would have in practice if we had access to a company's e-mail server, from where it would be possible to have access to the server's sending dates. Also regarding this same element, we must know when, in the case of informing times, certain time formats must be used, for example, with the time units written, as in 14 h and 30 min, instead of `14h30` or `14:30`. All content was written in Brazilian Portuguese, with the exception of foreign words that exist, besides area-specific vocabulary that needs to be used.

2.3. Sentence Structure

The text after a line break is a new email. This was done because of the data reading format for processing the tacit knowledge extraction, the layout of the dataset file, and to avoid problems with different line break characters in spreadsheet files for multiple operation systems. Therefore, if necessary, new periods of a sentence must be continued after a period "." and, besides indicating the end of a sentence, in any other situation we should use this sign in the file.

2.4. E-Mail Content Present in the Dataset

The format we tried to achieve is that, in each spreadsheet of the file, we had a folder of e-mails sent by a person. Therefore, the sending dates will not always be in sequence, as a person does not necessarily send an email every day. Nevertheless, the subject of the text is always inserted in the same context of professional activities, which is related to the position that the sender would have within the work environment of which he is a part.

Therefore, the texts can contain bosses texting their subordinates, as well as the opposite situation, that is, an analyst texting his superior. There may also be texts that are addressed to more than one person. This is the case of an individual who owns some form of business and is communicating with all of his customers. There may also be a person who is talking to someone who holds the same job title.

Finally, it should be noted that, in general, email texts are not usually very long. That is, most of the time we expect to find simple or compound sentences with two or three clauses, for example, contrary to what one would expect to find on a blog, in a report, or in a didactic text.

3. Methodology

In this section, we will describe the methodology that was followed to build the dataset. Initially, it was necessary to define a minimum sample size, which would guarantee the construction of a statistic to verify the adequacy of knowledge acquisition models. Furthermore, to ensure that the sensitivity of the acquisition models in recognizing labeled data had no bias, it was defined that an equal amount of each type of label was present in the sample.

Next, we had to ensure that the structure of the texts present in the sample reflected the environment in which they are inserted. After all, in our case, we are considering the

corporate world, but if we were to consider texts present in blogs or reports, the forms of writing would be different.

After that, the content of the texts was defined in such a way that we were left with a dataset where each set of emails had been written by a person in their daily work routine. We also wanted to ensure that some grammatical elements were present in some of the texts, so that they could later be identified during knowledge acquisition. Finally, we wanted to ensure that the dataset represented a set of texts with different forms of writing and with different professional experiences, without giving greater importance to certain subjects. Therefore, the calculated sample size was equally shared among the people available during the preparation of the emails. The following figure illustrates this entire process.

Figure 1 below shows that, after defining the amount of emails (Section 3.1), each content producer stayed responsible for creating a fraction of the total amount and followed the guidelines described in Section 3.4 to write the texts. And after defining the file structure (see Sections 2.1–2.3), the files were all consolidated in a format that enables the validation of a tacit knowledge extraction process.

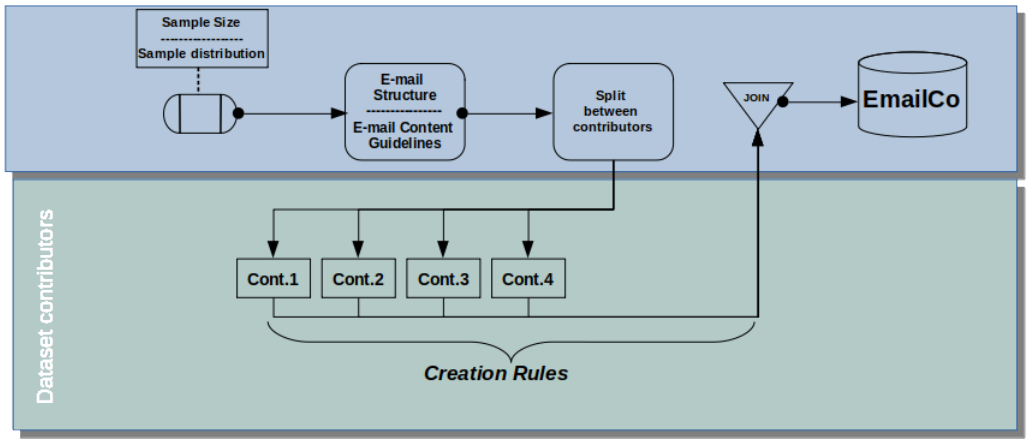

Figure 1. Methodology.

In the next subsections, first our objective will be to determine the initial size of the dataset (Section 3.1) and how the labeled data are distributed (Section 3.2). Next, we will detail the part of the contributors shown in the lower block of Figure 1, starting (Section 3.3) by defining the types of structures for building texts in emails that are generally the most commonly found in corporate environments. These formats were presented to the participants in the process of content creation, but its application was not strictly mandatory; after all, other formats are eventually found depending on the situation in which they are inserted.

Lastly (Section 3.4), the type of content that these structures should contain was presented, firstly the subject that they should deal with, then the grammatical elements that they should contain, and finally if the set of emails would have tacit knowledge embedded.

3.1. Sample Size

Considering that eMailMe supports the development of a labeled dataset that can be used for tacit knowledge acquisition, we started by defining its size. Such a definition is based on sampling techniques [11] and takes into account that the dataset should be composed of emails with and without tacit knowledge within them.

We built the confidence interval for the proportion of emails, which were identified for having some kind of tacit knowledge, and it would be correctly extracted. Through the confidence interval I, we can measure the magnitude of the error that we are committing when stating that the proportion of correctly extracted knowledge is a certain value $\hat{p}$, where

$$I = [\hat{p} - z(\gamma)\hat{e}, \hat{p} + z(\gamma)\hat{e}], where$$
$$\hat{e} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \quad (1)$$

In Equation (1), $\hat{p}$ is the proportion estimator, and $z(\gamma)$ is the normal curve index that delimits a confidence interval of size γ for a sample of size n.

For the proportion of a binomial distribution, we can estimate the size of our labeled dataset according to the tolerated error and the probability value of the interval containing the true value of the proportion of hits. Using the average of the proportion $\overline{p}$ as its estimator in Equation (1) and based on the fact that the binomial distribution can be approximated by the normal distribution, we have, for an error e, the estimated size

$$n = \left[\frac{z(\gamma)}{e}\right]^2 \overline{p}(1-\overline{p}) \quad (2)$$

Based on Equation (2) above, we established values to have a dataset of size n, considering a maximum error of 5% between the estimate $\bar{\mathbf{p}}$ and the real proportion of extracted emails with tacit knowledge, a confidence interval probability of 90%, and assuming that at least 55% of the tacit knowledge can be correctly identified. As a result, we calculate that we need a sample of 1630 texts simulating corporate emails.

We consider the dataset size estimate conservative because, due to available resources and established deadlines, we cannot commit to generating a larger dataset, and we are also assuming that the success rate of extractions is a tolerable minimum.

3.2. Sample Distribution

This dataset was designed to be used in a process of identification and extraction of tacit knowledge in a group of emails. To ensure that the initially created dataset had a uniform distribution between data labeled with tacit knowledge and those without, it was defined that half of the data were of one type and the other half of the other. Thus, it is possible to quantify the estimator for the proportion of correct answers in the case of identifying labeled emails and subsequently its confidence interval. Besides that, each spreadsheet should contain about 10 email texts, which led us to create about 163 spreadsheets.

3.3. E-Mail Structure

Just as Searle studied speech acts by focusing on the reason that all linguistic communications involve linguistic acts [12], in addition to defining different forms of speech acts, in our case we will also establish the types of considered communications. We need to direct the message formats that will be transmitted, considering that in our case we are referring to the communication that takes place through writing.

To elucidate the content of the texts that compose the base, we will detail four types of textual structures that were considered as possible to exist in corporate emails. And these structures were defined by contemplating the existing relationships in a corporation, for example, superiors talking to their subordinates or employees talking to their peers, and this can happen in different departments of a corporation or in the same sector. In addition, this professional communication can exist in a slightly different way in smaller companies, that is, those that do not have numerous departments, which may lead, depending on the case, to direct contacts with other companies or with their customers.

In communications between peers, many exchanges of information occur, which implies question and answer texts. For example, when a new computer system is being developed, many business area clarifications must be passed on to the developers. On the other hand, when we have a superior talking to subordinates, in general, texts with request formats are more common, such as when the project manager establishes to all involved

that the deadline for delivery of the system was advanced. However, there may also be requests between areas that do not have a hierarchy relationship, such as when a factory is requesting an input for one of its suppliers or when an announcement is being made, such as the delivery date of a shipment.

3.3.1. Notification

First, it is common for a specialist to notify others about something that is about to happen within their working group. As an example, we can cite *"The HR director is calling a meeting next Tuesday to introduce our new food service provider"*. So, we can see a grammatical structure composed of a subject, the action she is performing, why she is performing the action, and when a certain event will occur. Still thinking about notification emails, we can consider an example with the same grammatical elements as the previous one, but with a more commercial content, as in the case of a salesperson who could send an advertisement like, *"Our store will have a promotion to burn off our stock with discounts of up to 25% next week"*.

3.3.2. Answer

Then, we can cite the case where the specialist is answering something that has been asked. For example, *"Here's the file with the requested data correction. There was, however, no change in cost"*. In the example, in addition to the response about the change in cost, we also see something that is quite common in corporate emails, the sending of files or documents as an attachment to the email. Finally, the phrase describes what is being sent and the content of it.

Another answer format, which will be important for our experiment, is the case where the specialist adds an explanation of how something was done, as we see in *"The database was reprocessed considering the values of the last quarter for the revenue variable, because these values are now included in the current version of our policy"*. That is, in the example, we have the reason why a certain action was performed by the specialist.

3.3.3. Question

Naturally, the professional is answering something that was sent to him in the form of a question; this is another text format that must be considered. For example, *"How has your physical recovery period been after our last workout? Soon we will have to schedule a new evaluation"*. Here we have a direct question to the e-mail recipient about an activity that is performed by him with the sender's supervision, in addition to future planning, but without a defined date, for a future activity.

3.3.4. Requests

Another form of content considered is requests, where we could be referring, for instance, to requisitions for materials, as in the case of *"Would like to request 5 kg of Tuna to be delivered by next Friday"*. Here there is a hidden subject requesting a certain amount of material to be delivered within a certain deadline. Another example would be something more from the daily life of several offices, such as *"Please send the latest monthly report with the performance indicators of the wine campaign"*. Again there is a hidden subject with a description specifying what should be sent.

3.4. E-Mail Content Guidelines

Keeping the standards of Brazilian Portuguese, the subject contained in the text of an e-mail is free, with the exception of the following rules that must be respected.

3.4.1. Themes

Although the subject is free, it must be related to the corporation where the alleged specialist works and which was declared in the second line of the file. It must also be dealing with an activity in the area and be built in a similar way to what would be done in the corporate environment. Therefore, in this case, it will be necessary to take into account

the professional experience of who is writing the content or other forms of experience that he has had in his life, for example, contact with professionals in commercial areas, in the health area, or also in the service sector. In this way, we will have more foundation to build texts closer to what would occur in the daily communication of a specialist in their activity.

3.4.2. 5W1H Approach

Considering that tacit knowledge will be extracted based on the 5W1H approach, it is preferable to create sentences containing these elements grammatically; however, it is natural that not all sentences written by a person always contain all these components. Generally, an email sentence in a corporate environment always contains some action being reported, which would represent the "what". In fact, by definition, it is not possible to think about a sentence without having an action, which would make this the only mandatory element.

The other elements of 5W1H are the complements of this action, such as the location "where" it would be taking place and also "when" it would be taking place (the latter of which may or may not exist), as well as the "who", since there are sentences with the non-existent subject. Also, a "why" reason is not always present for certain decisions, much less who was responsible for them, for example, the information that a certain company will no longer provide an annual bonus to its employees, just like the "how", which tends to be more common in technical emails, where the professional needs to explain to another one how he arrived at a certain result. However, there may also be explanations without a grammatical structure with these characteristics, such as simply sending pertinent documentation that, by itself, could be explaining something to someone. Therefore, there are no exclusive restrictions on the presence of all these elements.

3.4.3. Tacit Knowledge

Finally, let us return to what was introduced in Section 2 about the simulated tacit knowledge embedded in the set of sentences and the process to extract them. We verified two ways of identifying this type of knowledge. The first one considers as tacit the content that is recurrent in the e-mail group; however, it is not repeated in the same text. In other words, as this knowledge has the characteristic of being intrinsic to the person, and therefore not explicitly exposed, the idea is that it would be present in the set of words that are repeated more frequently throughout the texts written by the specialist, eventually even in an unnoticed way.

To better exemplify this, consider an individual who works with a director. It is natural to suppose that part of the content written by that person contains the word "director" as the subject of the sentence, including sentences that contain observations learned through the coexistence that the person in question has experienced with the director. Therefore, this would consist of characteristics about the typical behavior of this director that were assimilated by this particular specialist over time. So, for this case, we must create texts with the same element as the subject of the sentence in a part of the emails of the set, and the rest would be without restrictions, with the exception, of course, of the other rules. Examples of this would be, *"Our director does not usually discuss such technical aspects in meetings, because…"*, *"This director does not make any approval without first…"*, and *"On 4 July at 3 h the director presented…"* among others phrases from different topics.

The second way seeks to identify the e-mail with words that are not recurrent in the rest of the e-mails set, in other words, something that does not specifically characterize the area of professional activity. However, if some tacit knowledge is present, as the text was written by the specialist about his professional activities, it may exist in the middle of this type of vocabulary. For this case, we created a sentence with a subject and the other grammatical elements (what he did, where, how, when, and why), with a group of words that were not repeated in other emails in the set. Figure 2 summarizes the guidelines described in this subsection for the e-mail creation process. After starting the creation process, we simultaneously have the processes that define the email theme and its

grammatical structure. The first takes into account the professional and personal history of the person, and the second the composition of the text with 5W1H elements. Finally, we must decide whether or not the email should have tacit knowledge embedded.

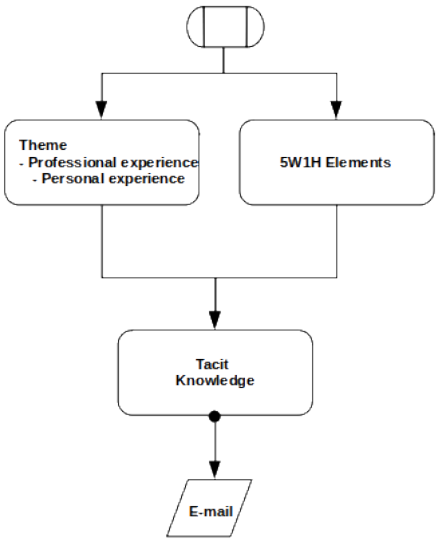

Figure 2. Creation rules.

3.5. Applying the Methodology

To ensure that the texts of the emails created have writing diversity, as well as content from real professional experiences from different areas, three people with different academic and professional backgrounds were hired to follow, along with one of the authors who has greater corporate experience, the methodology described in Section 3 and write 25% of the base each. The profiles of the hired specialists are as follows:

- a teacher of Brazilian Portuguese with 10 years of experience in K-12 courses;
- a person with more than 30 years of experience in laboring within the Civil Registry Office;
- a former staff leader of the Federal Regional Court, also with more than 30 years of experience in law issues.

In addition to the texts with content in their respective areas, such as *"Realmente, o comando da questão dois não está claro. Os alunos primeiramente devem interpretar o verso do poema para depois fazer a análise sintática. Vou reescrever do jeito que você sugeriu"* in the teaching area, *"Este ato é considerado irrevogável, pois, uma vez que o cidadão decidir a alteração, se arrepender após o deferimento do processo e quiser voltar como antes, não haverá essa possibilidade"* in the Civil Registry Office, and *"Assente jurisprudência no sentido de que proferida a Sentença, estando esta já assinada, e saindo o processo do Gabinete do Juiz, é dada como publicada, ainda que não esteja publicada pelos meios eletrônicos ou pela Imprensa"* in Federal Regional Court, the material was also created based on other situations they have experienced professionally, or on their respective life knowledge.

Although, in practice, corporate emails are not free from grammatical errors, all texts were revised so that there were no spelling, punctuation, or verb conjugation errors.

4. User Notes

As stated in Section 2.2, the choice for our file format was intended to create a table structure, as in a database, but in a file that does not depend on a management system to work with it. However, to deal with this file, our suggestion is to use the `read_excel` function from the Pandas library[1] for Python.

This function loads data into a Dataframe structure, making access to them more practical. If we assemble a tuple structure, we can in a simpler way manipulate the infor-

mation for other uses that the dataset may present. The tuple will be a pair of Dataframes, represented as an ordered pair, where the first position of the tuple contains a dataframe with the information lines about the group of emails, that is, the base identification, the area label, and the tacit knowledge label, and in the other position of the tuple, a Dataframe with the email IDs, the dates, and the text of the email.

$$< \underbrace{base\_id, area, knowledge}_{spreadsheet\ header}, \underbrace{email\_id, send\_date, email\_text}_{email\ content} >$$

Each of the worksheet tabs was named with the same base identification number, starting with 0 and going up to 163. This parameter must be informed to the `read_excel` function in the `sheet_name` option so that we can go through the entire file. In addition, we can inform the column names in the `names` parameter, which receives a list of values. And, finally, the lines that need to be skipped or read are stored in the `skiprows` and `nrows` parameters, respectively. In these cases, the first one also receives a list of values, and the second one receives an integer as its value.

To obtain a summary of the professional activity types, with content covered in the development of emails, it is possible to accumulate the values of the header field (`area`) in a list and use a summary function, such as `unique` from the Numpy library[2] for Python because the names used for this field were standardized.

However, there are alternatives to the use of Python and its libraries, such as the use of the statistical software R [13] and the `readODS` package[3], which, unlike the Python library, only reads ODS files. If you are working with XLSX files, you must use the `readxl` package[4]. Therefore, to load the dataset, we must call the `read_ods` function, since here, the parameter to skip the lines is called `skip`, and to inform the name of the worksheet tabs, we use the parameter `sheet`.

5. Conclusions

The dataset resulting from what was presented in this article produces simulated sets of corporate emails labeled by the presence or absence of embedded tacit knowledge. Such emails were constructed taking into account the 5W1H approach for knowledge extraction and are intended to validate a tacit knowledge extraction process. Nevertheless, it can also be used for training neural network models, since we have groups of emails labeled by area of activity in which they are inserted, or testing text mining algorithms. After everything, at the end of the process, 1660 texts were obtained with a grammatical structure that, in general, is used in emails, dealing with different areas of professional activity and respecting the grammar rules of the Brazilian Portuguese language. It should also be noted that, to the best of our knowledge, this is the first dataset on this theme and whose texts are grammatically correct. To ensure this, the texts have been double-checked, in addition to the fact that one of the members of the group, who participated in the creation of the emails, is a Brazilian Portuguese grammar teacher.

However, an existing limitation in this work was the number of qualified people to assist in the creation of dataset content. This was due to time issues in the project schedule for finding people, in addition to instructing them according to the methodology presented here. With more collaborators, we would have a more diversified content dataset in terms of professional experiences, and it would also be possible to structure a statistical process for verifying the quality of the created content to produce an additional guarantee that it reflects real corporate situations. By hiring a significant number of people with corporate experience, we could ask them to read and mark each of the dataset texts as coming from a corporate email or not. Then, we would have to build a metric and the appropriate statistics to check the percentage of texts that would be identified as coming from an employee in a work environment.

A possibility for further research which would expand the scope of this work would be the identification of new ways of embedding tacit knowledge in texts. However, this

subject would require further textual analysis that could not be included in the scope of this project.

Furthermore, the structure used in the file allows for different ways of grouping data, such as creating a single group with all the texts or grouping them by area. Finally, although it was specifically built with the insertion of simulated tacit knowledge, it can have other uses than extracting this knowledge, since when reading some of these excerpts separately, it is not possible to identify distinct characteristics of any other common text. This can be viewed in the following example, which contains the knowledge, *"A nuvem fica melhor fotografada se utilizarmos um filtro de polarização sobre a lente da camera"*, compared to *"Gostaria de saber qual é a data máxima para o lançamento das notas do segundo trimestre. Procurei no calendário, porém essa informação não consta lá"*, which does not.

Author Contributions: Conceptualization, A.A.d.M.G.U. and A.A.F.B.; methodology, A.A.d.M.G.U. and A.A.F.B.; software, A.A.d.M.G.U. and A.A.F.B.; validation, A.A.d.M.G.U. and A.A.F.B.; formal analysis, A.A.d.M.G.U. and A.A.F.B.; investigation, A.A.d.M.G.U. and A.A.F.B.; resources, A.A.d.M.G.U. and A.A.F.B.; data curation, A.A.d.M.G.U. and A.A.F.B.; writing—original draft preparation, A.A.d.M.G.U. and A.A.F.B.; writing—review and editing, A.A.d.M.G.U. and A.A.F.B.; visualization, A.A.d.M.G.U. and A.A.F.B.; supervision, A.A.d.M.G.U. and A.A.F.B.; project administration, A.A.d.M.G.U. and A.A.F.B.; funding acquisition, A.A.d.M.G.U. and A.A.F.B. All authors have read and agreed to the published version of the manuscript.

Funding: This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brazil (CAPES)—Finance Code 001.

Data Availability Statement: At the end of the entire process, the dataset was made available on the open platform Github at https://github.com/ariceak/emailCo (accessed on 25 July 2023) so that the data can be reused in other studies or even expanded.

Acknowledgments: The development of the dataset presented in this article was only possible thanks to the contribution and dedication of former federal civil servant Maria B. de M. Galvão and civil registry clerk Janete Lopes, who used their knowledge and experiences of these environments to compose the texts, leaving them with a content closer to reality, as well as the Portuguese teacher Beatriz M. Bezerra, who, in addition to providing her experience in the teaching area, also helped in the discussions about the grammatical rules used. For all this, we would like to express here our most sincere thanks to the work of these people.

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

Abbreviations

The following abbreviations are used in this manuscript:

KM	Knowledge Management
5W1H	Who, What, When, Where, Why, and How

Notes

1 https://pandas.pydata.org/ (accessed on 25 July 2023)
2 https://numpy.org/ (accessed on 25 July 2023)
3 https://cran.r-project.org/web/packages/readODS/index.html (accessed on 25 July 2023)
4 https://cran.r-project.org/web/packages/readxl/index.html (accessed on 25 July 2023)

References

1. Jurisica, I.; Mylopoulos, J.; Yu, E. Using Ontologies for Knowledge Management: An Information Systems Perspective. In Proceedings of the Annual Conference of the American Society for Information Science, Washington DC, USA, 1–4 November 1999.
2. Mohammad, A.H.; Al Saiyd, N.A.M. Guidelines for Tacit Knowledge Acquisition. *J. Theor. Appl. Inf. Technol.* **2012**, *38*, 110.
3. Hamborg, F.; Breitinger, C.; Gipp, B. GiveMe5W1H: A universal system for extracting main events from news articles. In Proceedings of the INRA-International Workshop on News Recommendation and Analytics, Copenhagen, Denmark, 19 September 2019.
4. Supnitchaisiri, M.; Natakuatoong, O.; Sinthupinyo, S. The innovative model for extracting tacit knowledge in organisations. *Int. J. Knowl. Manag. Stud.* **2020**, *11*, 81–101. [CrossRef]

5. Carnaz, G.; Nogueira, V.; Antunes, M. A Graph Database Representation of Portuguese Criminal-Related Documents. *Informatics* **2021**, *8*, 37. [CrossRef]
6. Carnaz, G.; Antunes, M.; Nogueira, V.B. An Annotated Corpus of Crime-Related Portuguese Documents for NLP and Machine Learning Processing. *Data* **2021**, *6*, 71. [CrossRef]
7. Islam, M.T.; Hasan, K.M.A.; Hossen, M.I. Classification and Resource Generation for Bangla Emails Based on Machine Learning Algorithms. In Proceedings of the 2022 25th International Conference on Computer and Information Technology, ICCIT 2022, Cox's Bazar, Bangladesh, 17–19 December 2022.
8. Cha, I.; Oh, J.; Park, C.Y.; Han, J.; Lee, H. The Grind for Good Data: Understanding ML Practitioners' Struggles and Aspirations in Making Good Data . *arXiv* **2022**, arXiv:2211.14981.
9. Hristov, E.; Petrova-Antonova, D.; Petrov, A.; Borukova, M.; Shirinyan, E. Remote Sensing Data Preparation for Recognition and Classification of Building Roofs. *Data* **2023**, *8*, 80. [CrossRef]
10. Alshammari, T.; Alshammari, N.; Sedky, M.; Howard, C. SIMADL: Simulated Activities of Daily Living Dataset. *Data* **2018**, *3*, 11. [CrossRef]
11. Bussab, W.O.; Morettin, P.A. *Estatística Básica*; Saraiva: São Paulo, Brazil, 2006; p. 307.
12. Searle, J. *Speech Acts: An Essay in the Philosophy of Language*; Cambridge University Press: London, UK, 1969; p. 16.
13. The R Foundation. R. Version 3.6.3. 2020. Available online: https://www.r-project.org/ (accessed on 25 July 2023).