Text Analysis Based Human Resource Productivity Profiling

Basudev Pradhan(✉), Siddharth Swarup Rautaray, Amiya Ranjan Panda, and Manjusha Pandey

School of Computer Engineering, KIIT Deemed to be University, Bhubaneswar, Odisha, India
{basudev.pradhan,siddharthfcs,amiya.pandafcs, manjushafcs}@kiit.ac.in

Abstract. Email being an efficient, cost-effective, real-time communication mode results into effective productivity among the professional in the organization. It constitutes almost 90% of daily office procedures in organizations, hence the productivity of organizations depends heavily on the text communicated in emails. The presented research work focuses on email profiling in organizations based on mail text interpretation and analysis. In the proposed work we will be working on datasets containing email communication of ENRON Corporation as test case. The profiling would be done using Text interpretation and analysis algorithm using machine learning algorithms. The BoW will be implemented to analyze and predict the characteristics of incoming and outgoing emails, then these could be mapped and profiled as per the behavior of employees into 3 categories of productive based on positive responses, neutral and non-productive based on negative responses.

Keywords: Email profiling · text interpretation and analysis · ENRON dataset · machine learning · Bag of Words

1 Introduction

Professionals in an Organization may be profiled based on the analysis of text used by them in the email communications while performing their role for contribution to organization's goal, as almost 90% of the communication in the current age of digital data transformation is in email format [1]. The characteristics of words in email communication indicates about the behavior and attitude of professional and is a reflection of how effectively the professional works. Many research and existing labor data indicates the professional using positive terms in their email communication work effectively and efficiently and tend to make the work environment more favorable for their co-workers.

The presented research work focuses on identification and categorization of choice of words used by different professionals of the organization in their email communication. The email profiling would be done based on mail text analytics using BoW machine learning algorithm. The input data taken from ENRON dataset will be preprocessed for the removal of email text formalities and the information is integrated in a weighted

P. Pareek et al. (Eds.): IC4S 2023, LNICST 536, pp. 254–262, 2024.
https://doi.org/10.1007/978-3-031-48888-7_21

manner based on the number of positive, negative and neutral words used in the email text. The expected output will be laid down from natural language processing and narrative analysis for email profiling. We will use BoW algorithm to classify/categorize emails as an assessment tool for the productivity of the employees in the organization. This email profiling based on productivity employees can be done on weekly or hourly bases and attempts can be made to develop trend models for employee productivity analysis.

2 Related Work

Dr. Deborah Fallows [2] in his research work has deduced email to be most effective way of communication over physical meeting & telecommunication citing a percentage of 63%. Dr. Fallows suggested the employees used emailing 67% of time for professional communication 26% of time for personal communication and 15% of for gossips. Thus, formalizing email as a most preferred communication mode for official purposes. Another research work done by Emmanuel Gbenga Dada [3] highlighted efforts made by different researchers to solve the spam and ham problem using effective classifiers of his review of machine learning algorithms for identification of spam and ham messages. Also, a comprehensive review for the emails spam classification problem has been done by Mansoor Raja [4]. His findings are more focused on usage of specific algorithm namely Naive Bayes and SVM. The author has also suggested usage of multi algorithm-based system over single algorithm system. K. Thirumoorthy [5] has proposed a method based on term frequency distribution measure (TFDM) for investigation of performance of two of the most preferred algorithm Naive Bayes (NB) and Support Vector Machine (SVM) over of bench mark text corpus. He suggested a thorough experimental analysis depicts better performance of TFDM over various well-known filter techniques (DF, ACC2, IG, MI, DFS, NDM and TRDL). Another researcher Maryam Hina [6] in her study of multilevel email classification has generated promising linear regression accuracy along with a comparison with logistic regression. She suggested that logistic regression provided 91.9% accuracy and also included incorporation of block chain for storage and access of analyzed data as a future work. A lot of work for classification of spam and ham mail has been done using content base features instead of behavioral features that generates the requirement of multi-folder classification of emails considering heterogeneous details of emails rather than content only. This conclusion by Namrata Shroff [7] and her research group emphasizes more on multi-folder classification of emails and generation of context based on the content of emails. The same exclusion of context is evident from all research literature available [8–15] for analysis of emails, our work thus emphasizes on generation of context of employee productivity by analyzing the text used in the content of email.

3 About ENRON Dataset

This proposed work is based on ENRON data set that was collected and prepared by CALO problem, which is cognitive assistant that Learns and Organizes. The dataset has made public for academic and research purposes related to email profiling text analytics

and other related research domains. The ENRON dataset contains email communication samples of about 150 users of the ENRON corporation which amounts to about 0.5 million email exchanges. The email exchanges included in the dataset are mostly senior management professionals for ENRON corporation. This data was made public by hosting it on the web by FERC (Federal Energy Regulatory Commission) during its investigation on reasons for failure of ENRON Corporation. The dataset features text and other contents which contains only the text communication among different users of the ENRON corporation. This makes the ENRON datasets most suitable for our research work based on email profiling using text analysis and interpretation. Our proposed method to work on this ENRON dataset consists of 3-phases:

Phase-1: Text and Content extraction from email corpus

Phase-2: Text interpretation and analysis

Phase-3: Email profiling and categorization/classification using BoW algorithm

Problem Statement:

The objective of projected research work would be 3-fold defining our problem statements as follows:

Obj-1: Feature extraction of text and content extracted from email corpus.

Obj-2: Email profiling based on mail text interpretation and analysis.

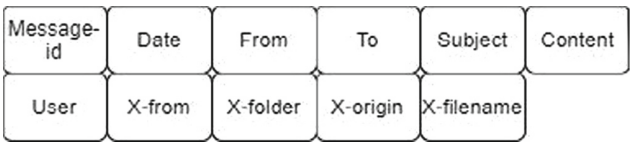
Obj-3: Categorization/classification of professionals based on email profiling using BoW algorithm.

4 Proposed Approach

For our research, we have developed the approach as depicted by Fig. 1. The proposed approach has been designed as the flow between the extractions of mail content from the ENRON email dataset to employee categorization based on their productivity mapped to the frequency of type of words used by them in their mail contents.

The Enron email data set has around 517,431 digital communications. This dataset has been provided by the FERC (Federal Energy Regulatory Commission) for academic and research purposes; it has data of 150 users in 3500 different folders available for analysis. The same dataset has been used in this research work with following features.

1. Message-ID,
2. Date,
3. From,
4. To,
5. Subject,
6. User, X-from,
7. X- folder,
8. X-origin, X-filename.



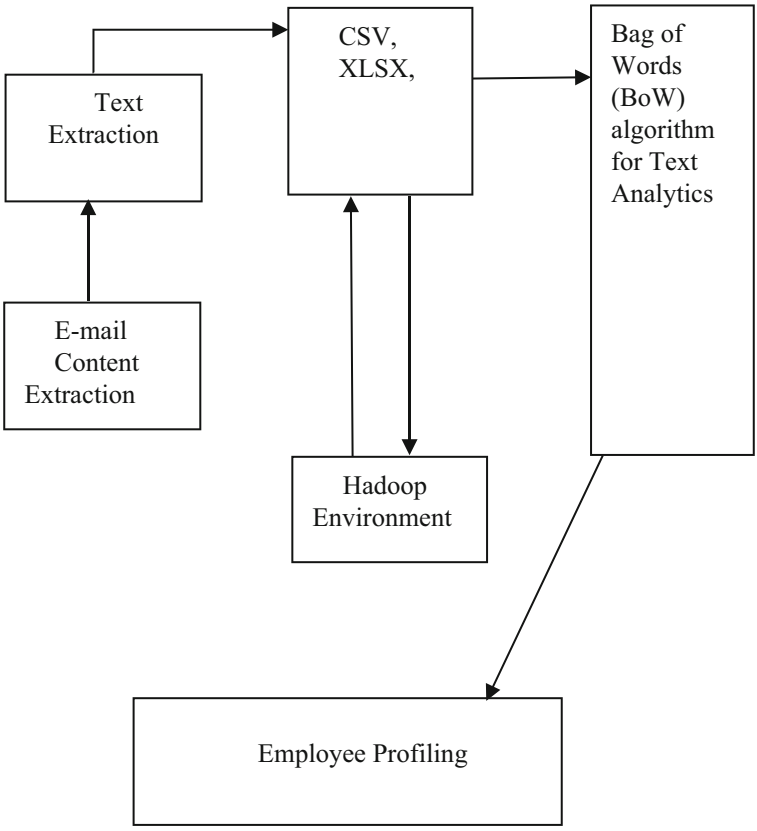


Fig. 1. Proposed Approach

The data was preprocesses in order to build a model to classify the emails responses into three categories of positive, negative and neutral. After this, the data set can be used for exploratory data analysis and visualized.

5 Feature Extraction of Emails

The data preprocessing based on the feature accepted by the classifier model takes text from the emails just including subject and body of an email; The data is filtered based on: 1) Removal of stop words like "Re:", "Fw:" and "Fwd:"; 2) The data is refined for stop words like "From" and "To".

For further classification and analysis specific conditions like emails with no subjects or "No Subject" or emails with no body are removed from the analysis candidate set.

The identification of text features after preprocessing of the content extracted from the mail mitigating all the discrepancies of email and in turn text categorization into set of positive and negative set of words using BOW model. The individual categorization of employees is further mapped to the output of the BOW model for individual employee

as extracted and stored from data set as a separate array. Thus, the mail content still remains the vital feature of email categorization which in turn helps in the employee categorization based on email categorization. The subjects and contents of the emails have not been included in the feature set as they are redundant many times with very less relevance to the email content while the word frequency words of body are graded for high frequency of words and not graded for lesser frequency of the words to enhance the text analytics based on the BOW model.

6 BoW Model

Bag of Words (BoW) model is a technique for extracting the features from text data. It is a method that is frequently used to describe the meaning of a document. It produces a fixed-length vector based on the frequency of words/terms. A term that is used frequently suggests that the document has more to do with that term and should therefore be given a higher value than the other terms. This is done by creating a Term Frequency table, which counts the number of times each term appears in the document. Based on the frequency of each term inside the content, the Term Frequency creates a vector space model of the document.

6.1 Data Extraction

Data extraction is the process of collecting the data from a variety of sources. This is the initial step for any type of text analysis.

6.2 Data Pre-processing

Data Pre-processing is the procedure of transforming the data in appropriate format, so that the data can be used efficiently according to the model. Here some of the steps for data pre-processing:

- Remove the stop words (most common words like ‘the’, ‘is’, ‘a’ and etc.)
- Convert the texts of the document in lower case as the case has no meaning.
- Remove punctuations and special characters.
- Make a list of all the words in our model vocabulary.
- Count the frequency of words in our text document.

6.3 Bag of Words Model

The Bag of Words model may be binary Bag of Words or Bag of Words. In both cases we have to construct a vector to indicate the appearance of word. For Binary Bag of Words, the vector contains either 1 if the word is present or 0 if the word is not present. And for the Bag of Words, the vector contains the value as per the frequency of the word in a sentence.

Example: (each and every sentence from different mail)

Let we want to vectorize the following:

S-1: The information in this email may be confidential and/or privileged.
S-2: We are being billed for this service and I do not know who is using it.
S-3: This email has the details of the service.

After removing the stop words, punctuations and lower-case conversion the sentences will be as follows:

S-1: information email confidential privileged
S-2: billed service know using
S-3: email details service

Now, we have to count the frequency of word in a sentence and we have to vectorize our document.

Docs	information	email	confidential	privileged	billed	service	know	using	details
S-1	1	1	1	1	0	0	0	0	0
S-2	0	0	0	0	1	1	1	1	0
S-3	0	1	0	0	0	1	0	0	1

Hence, the resultant vectors are:

$$S-1 : [1, 1, 1, 1, 0, 0, 0, 0, 0]$$
$$S-2 : [0, 0, 0, 0, 1, 1, 1, 1, 0]$$
$$S-3 : [0, 1, 0, 0, 0, 1, 0, 0, 1]$$

We observe that, in Bag of Words methodology, we refined source data into dataset which lose contextual information, only lists vocabulary data with frequency. Hence, it is very helpful in machine learning technique as the large volume of data is processed.

The experimental setup was established for sample data generated through extraction of 1000 users' data from the Enron data set. For training and testing purposes the data was divided into 70–30 ratio. Where 70% was utilized for training of the generated model and 30% was utilized for testing. The BOG Model generated and implemented through python coding in jupyter note book utilizing the numpy, pandas, re, nltk, seaborn, matplotlib etc.

7 Experimental Result

This paper has analyzed year-wised emails for positive, negative and neutral words. It has traced and analyzed the emails of particular years as 1999, 2000 and 2001.It has also conducted separate analysis study of 1000 emails received by Enron Corporation for a tenure of 5 years.

The following results have been generated after the classification of the pre-processed dataset. The following Fig. 2 of classification of digital data communication for all 12 months of the year 1999 depicts more negative responses during the mid of the year and lesser during the start of the year which is the duration of lesser work pressures.

The following Fig. 3 of classification of digital data communication for all 12 months of the year 2000 depicts more negative responses during the mid of the year and again lesser during the start of the year which is the duration of lesser work pressures.

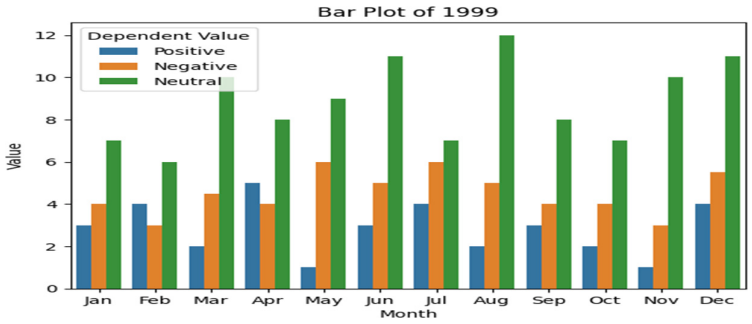


Fig. 2. Bar Graph of emails analysis of year 1999

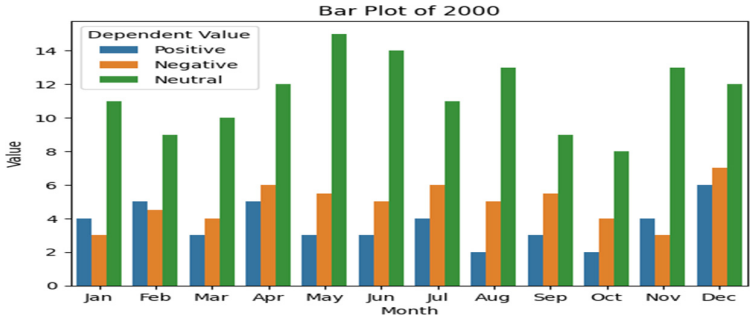


Fig. 3. Bar Graph of emails analysis of year 2000

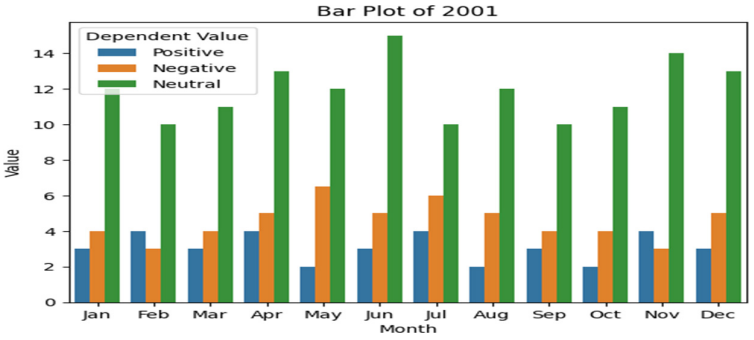


Fig. 4. Bar Graph of emails analysis of year 2001

The above Fig. 4 of classification of digital data communication for all 12 months of the year 2001 again depicts more negative responses during the mid of the year which also confirms the efficiency of our machine learning model for further utilizations also.

The results of classification of text in the dataset for 5 years depicts the increasing of negative responses year by year as is shown in Fig. 5.

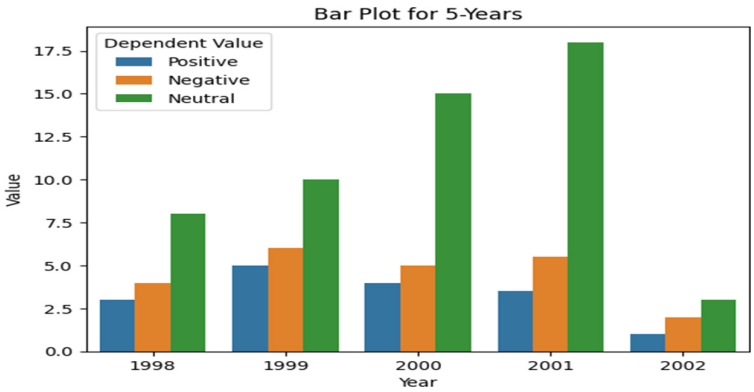


Fig. 5. Bar Graph of emails analysis of 5-years

8 Conclusion

After reassuring the efficiency of our proposed model for the data set of three different years we performed the classification of digital data communication for all 5 years of data available in the data set and the results were depicting the increase in negative responses year by year which is evident by the shutting up of the company after the year 2002. The results depict in Fig. 5 emphasize correlation between the productivity of the employees to the usage of positive, negative and neutral words in their day-to-day communications. Thus, the proposed research work can be utilized by the organization for establishment of productivity quotients of their employees based on the classification of text chosen and utilized by them during their day-to-day digital communications.

References

1. Aufreiter, N., Boudet, J., Weng, V.: Why marketers should keep sending you e-mails. McKinsey & Company (2014)
2. Fallows, D.: Email at Work. Pew Internet & American Life Project (2002)
3. Dada, E.G., Bassi, J.S., Chiroma, H., Adetunmbi, A.O., Ajibuwa, O.E.: Machine learning for email spam filtering: review, approaches and open research problems. Heliyon **5**(6), e01802 (2019)
4. Mansoor, R.A.Z.A., Jayasinghe, N.D., Muslam, M.M.A.: A comprehensive review on email spam classification using machine learning algorithms. In: 2021 International Conference on Information Networking (ICOIN), pp. 327–332. IEEE (2021)
5. Thirumoorthy, K., Muneeswaran, K.: Feature selection for text classification using machine learning approaches. Natl. Acad. Sci. Lett. **45**, 51–56 (2021). https://doi.org/10.1007/s40009-021-01043-0
6. Hina, M., Ali, M., Javed, A.R., Srivastava, G., Gadekallu, T.R., Jalil, Z.: Email classification and forensics analysis using machine learning. In: 2021 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/IOP/SCI), pp. 630–635. IEEE (2021)

7. Shroff, N., Sinhgala, A.: Email classification techniques—a review. In: Kotecha, K., Piuri, V., Shah, H.N., Patel, R. (eds.) Data Science and Intelligent Applications. LNDECT, vol. 52, pp. 181–189. Springer, Singapore (2021). https://doi.org/10.1007/978-981-15-4474-3_21
8. Nandhini, S., KS, J.M.: Performance evaluation of machine learning algorithms for email spam detection. In: 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic- ETITE), pp. 1–4. IEEE (2020)
9. Das, N., Shankar, S., Dash, B., Mohan, J., Pandey, M., Rautaray, S.S.: Productivity profiling of organizations based upon communication interpretation and analysis. In: 2021 5th International Conference on Information Systems and Computer Networks (ISCON), pp. 1–6. IEEE (2021)
10. Noever, D.: The enron corpus: where the email bodies are buried? (2020)
11. Harrison, J., et al.: Quantifying use and abuse of personal information. In: 2021 IEEE International Conference on Intelligence and Security Informatics (ISI), pp. 1–6. IEEE (2021)
12. Negangard, E., Fay, R.: Electronic Discovery (eDiscovery): Performing the Early Stages of the Enron investigation. Issues Acc. Educ. 35 (2019). https://doi.org/10.2308/issues-16-064
13. Ali, R.S., Gayar, N.E.: Sentiment analysis using unlabeled Email data. In: 2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE), pp. 328–333 (2019)
14. Jacob, I.J.: Performance evaluation of caps-net based multitask learning architecture for text classification. J. Artif. Intell. **2**(01), 1–10 (2020)
15. Karim, A., Azam, S., Shanmugam, B., Kannoorpatti, K., Alazab, M.: A comprehensive survey for intelligent spam email detection. IEEE Access **7**, 168261–168295 (2019)