# An Anlysis on Email Classification by Using Topic Modeling

# An Anlysis on Email Classification by Using Topic Modeling

Md Shafayat Hossain
Applied Informatics, Graduate School of
Science and and Engineering
Hosei University
Supervisor: Akihiro Fujii

***Abstract*—Email Classification is a broad term. It might be used to separate spam email, detect phishing emails, detect threat emails, and many other aspects. Many papers have been published on these topics over the years. This paper is focusing on classifying email into different categories by using the topic modeling technique. For that, the email body will be analyzed to categorize emails from scratch. In topic modeling, documents are considered a collection of topics, and topics are considered a collection of words, also known as a bag of words. We used the Latent Dirichlet Allocation topic model, also known as the LDA topic model to extract topics from email dataset. We used the Enron email dataset for our experiment, which is the largest open-source email dataset. The expected outcome will be human interpretable topics that can easily be identified and categorized by labeling them.**

***Keywords—email, classification, Enron, LDA, machine learning***

## I. Introduction

Email is an electric form of letter. Electronic mail, also known as email, has become an indispensable part of our daily life. Specially, business activities in nearly all aspects of commerce. The main reason for this is accessibility. Nowadays, anyone who owns a smartphone has their email address to use along with it. Over the years email user increased significantly. According to a statistical analysis by *The Radicati Group* more than 3.9 billion people have their own email addresses and it will be 4.37 billion approximately at the end of 2023. They also assume that 300 billion emails have been trafficking every day. These numbers are huge and give us an idea about the importance of email management. With the increasing number of email users, email communication is also getting popular day by day. Nowadays, we often see an email-based survey, marketing, alerts, notification, and other services. Within the organization, every formal and business communication is dealt with email. And almost every organization and agency received tons of email every day regarding their services, complaints, feedback, and so on. To understand the nature of the queries and the feedback through email, either they have to spend resources and more time to evaluate the massive number of emails, or it can be sorted out smartly within a feasible time with fewer resources. Moreover, for that reason, email classification is getting more focus.

Over the years, there is a lot of research to classify and retrieve information from this enormous amount of email datasets. To manage emails, the email classification terms get introduced. Email classification is a broad field. It could be detecting spam mail, separating phishing mail, detecting threat mail, and many other aspects. Here, we are focusing on email classification based on their topic extraction using the topic modeling method. We will use the Latent Dirichlet Allocation, also known as the LDA model, which is a generative statistical model.

## II. Email Classification Methodology

### A. Different Aspects

There are lot of area of email classification. A survey paper of IEEE [2] indicates at least 15 classes of areas of email classification has been explored. The paper minimizes it to major five areas and those are spam email classification, phishing email classification, spam and phishing email classification, Multiple categories classification and others classification which includes terrorist email classification, image-based spam classification etc. This paper will only focus on multiple categories classification. According to that IEEE survey paper[2] based on email classification from 2006 to 2016, 98 papers were published regarding email classification. The paper considered only the research which is conducted for the English language-based email dataset. Sixty four percent of the total number of released papers during that period is either spam or phishing email classification related. And Twenty one percent (20 papers) of 98 papers were in multi categorization related.

### B. Previous Used Method and Technology

According to the survey paper, the widely used method while using machine learning classifier is the supervised learning. Most of the research paper uses SVM classifier. But in this paper unsupervised learning will be applied to determine the classification. Because, the feature to classify the email will only be the email body. And the email dataset is not labeled either. The unsupervised method will explore the dataset as it is, and it will decide the topic classification without any prior forced absolution. This way no topic will be ignored.

### C. Dataset

For this research, Enron email dataset has been used. It is the largest open source dataset available online. It contains approximately a half million emails generated by employees of the Enron Corporation. It was obtained by the Federal Energy Regulatory Commission during its investigation of Enron's collapse. It is the only mass collection of real data available for researcher. That is why for this research it has

been chosen. This dataset contains email from 158 unique users. As the number of emails is huge, a subset of twenty thousand email will be used for the research purpose.

### D. Topic Modeling and LDA

The use of topic modeling is not new. However, there are remarkably few papers utilizing the method for categorizing the email dataset. Topic modeling is one of the most powerful techniques for data mining, latent data discovery, and finding relationships among data, text documents. It is a branch of unsupervised natural language processing that is used to represent a text document with the help of several topics that can best explain the underlying information in a document.

Latent Dirichlet Allocation also known as LDA is generative statistical model that allows a combination of observation to be explained by an unsupervised group. The concept is similar to K-means clustering, but the real difference is clustering does not allow one cluster member to be included into another cluster. But, in LDA a member can be belongs to multiple groups. This groups are known as topic. That is why it is also known as topic modelling. For unsupervised email classification, LDA has been considered. The main reason is the way it approaches to define a groups or topics and allows a word or component to be a part of multiple groups. And it has a proven record to a good classifier in text-based article classification by extracting topics and it is also used in bioinformatics area also. Furthermore, the field of topic modeling has not been explored well enough. In 2003, David Blei, Andrew Ng, and Michael Jordan introduced Latent Dirichlet Allocation in a research paper. It is described as a "generative probabilistic model of a corpus." It is widely used for performing Topic Modeling — a statistical technique that can extract underlying topics from a corpus. Let us have a look at the LDA topic model.

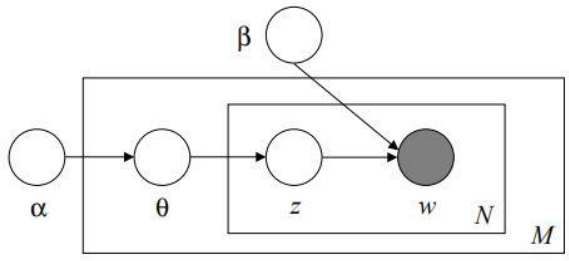


Figure 1: LDA Model

Here, M is number of documents, N denotes number of words in each document, *w* is a word in a document. And ***w*** (bold *w*) represents a document (i.e. it is the vector of *w*'s) of *N* number of words, z is a single topic from a set of *k* number of topics. Θ is the distribution of topics, one for each document. α is a parameter vector for each document (document — Topic distribution). β is a parameter vector for each topic (topic — word distribution).

LDA works like the following steps:

- It randomly assigns each word from each document to one of the $K$(given) topics.
- For each document $d$, It will assume all the topic assignments are correct except the running one. Then it calculates two proportions of probability. They are,
  - Words in document $d$ which is currently assigned to topic $t = p(t|d)$
  - Assignment of topic $t$ in all documents which come from this word $w = p(w|d)$
- And finally, it will reach a state where topic assignment make sense.

### E. Topic Coherence

Topic coherence score is trying to quantify the semantic similarities of the high scoring words within each topic. A high score means the result is more human-interpretable. So, naturally higher coherence score means a better model. The score is determined by the following equation.

$$Coherence = \sum_{i<j} score(w_i, w_j) \quad (1)$$

Here the pair of w represent the pair of weights of UCI measures and UMass measures. UCI measure is based on a sliding window and the pointwise mutual information of all word pairs of the given top words. UMass is based on document cooccurrence counts, a one-preceding segmentation and a logarithmic conditional probability as confirmation measure.

## III. Research Methodology and Result Analysis

### A. Data Implementaion and Preprocessing

For Experiment Jupiter notebook has been used. Pandas has been used for data implementation. And gensim library used for model implementation. Initially, from raw data only email body, sender and receiver is extracted for research purpose. Empty body, sender ore receiver emails have been dumped. For data cleansing following steps are followed:

- First, using regular expression we remove punctuation and junk word. Then we lowercased the whole email body to avoid case sensitivity.
- Secondly, sentences are broken into words and added to a list. Which is also known as tokenization.
- Then, we remove the stop words by using Nltk libraries. Now the corpus is ready to put into the model for analysis.
- We use N-gram to concatenate words that are meant to be together. Then performed the lemmatization which convert a word into its original state. For example: Going, Goes, Gone. These words lemma will be Go.

- To build our LDA model, we need a word2id dictionary. In a word2id dictionary, every word is mapped by a unique id. We use the gensim library to create this word2id dictionary. For example: {'forecast': 0, 'boat': 1, 'business': 2…}. Here each word is mapped to a unique id to represents themselves in the Dictionary.

- In the corpus, every word is represented by its id which is paired with the number of occurrences in a particular document. For example: [(13,1), (18, 5)] means the first-word id in the corpus is 13, and it occurs only once the current document. And the second-word id is 18, which occurs five times in the current documents.

### *B. Model Implementation*

For LDA model implementation genism API is used. So, only the require parameter should be determined to make it work. LDA has one flaw that it needs human interpretation. For creating the model one must have to decide how many topics or categories would be in the model. The other parameters are the corpus, the dictionary, which is numeric representation of words, number of documents in each training chunk, number of passes which determines how many complete passes needed for the training. And the hyperparameter alpha and beta; both values are ideally less than 1. We consider a range of value in between 0 to 1 and finally settled with 0.1 for alpha and 0.01 for beta which works well for our dataset.

For this experiment number of topics is considered initially 10. Then we have tried a range of values in between 2 to 35. After observing the outcome, we set our topic number to 20. The outcome is twenty topics. And each topic contains ten words with higher weights. Weights determine the word contributions in a topic. Here is a visualization.

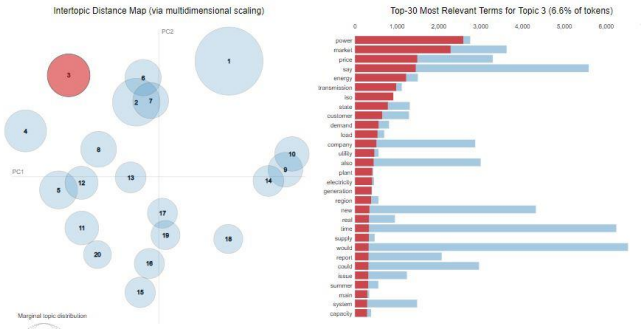


Figure 2: Topic Distribution

Here, the circle represents a topic. The bigger the circle is the topics influence is greater. The right side is showing the topics word. Here is look on our topic 20 topics with their highest contribution calculated based on words weight value, in a single topic:

| Topic_Perc_Contrib | Keywords |
|---|---|
| 0.9488 | game, week, play, last, team, say, start, year, allow, look |
| 0.8272 | order, offer, value, company, share, pay, would, must, may, number |
| 0.7552 | image, free, travel, day, special, offer, visit, click, holiday, save |
| 0.9299 | way, buy, may, purchase, available, click, time, save, travel, information |
| 0.8226 | business, service, cost, include, account, agreement, continue, statement, fee, would |
| 0.8913 | year, gas, trading, price, say, weather, trade, month, government, trader |
| 0.5984 | know, people, take, let, make, want, good, man, friend, hope |
| 0.5742 | let, know, forward, need, want, look, would, attend, interested, go |
| 0.7083 | message, original, deal, enronxgate, position, change, show, know, bid, value |
| 0.6761 | number, trade, call, deal, thank, need, volume, confirm, enter, day |
| 0.7257 | say, make, name, year, help, company, take, world, group, also |
| 0.9501 | font, size, align, right, leave, tr, table, td, color, width |
| 0.7326 | new, report, stock, free, access, page, market, time, security, internet |
| 0.7308 | need, would, pm, site, call, forward, week, question, comment, list |
| 0.6155 | enroncom, email, contact, web, deanlaurent, system, available, offer, ericbass, contain |
| 0.6371 | go, know, think, see, next, week, time, guy, back, would |
| 0.6000 | request, information, schedule, question, meeting, provide, employee, follow, process, change |
| 0.8195 | power, market, price, say, energy, transmission, iso, state, customer, demand |
| 0.5435 | go, get, would, see, think, want, day, good, come, thing |
| 0.7828 | mail, message, original, receive, send, may, contact, review, basis, party |

Figure 3: Topics with their Highest Contribution Value over a Singe Document(Email Body)

Looking at the topics, ten to twelve topics can easily be interpretable. Although the expected outcome is higher due to some limitation, we accept this result and will discuss it later.

### *C. Result Analysiss*

For evaluation, we use a coherence score. A higher coherence score means the topic is more interpretable. LDA offers a lot of parameter tuning like a topic number, alpha, beta, number of passes, etc. First, tried with the default settings predefined by the gensim library, then we tweaked alpha and beta parameters to get a good result. But, for a better outcome, we need to change the topic number value as well. That is why we created a function that iterates the topic value within a given condition, and it also shows the coherence value for the different number of topic selection. The following figure shows the before and after changing the alpha value. Although we have tried several values, we consider the best outcome only.

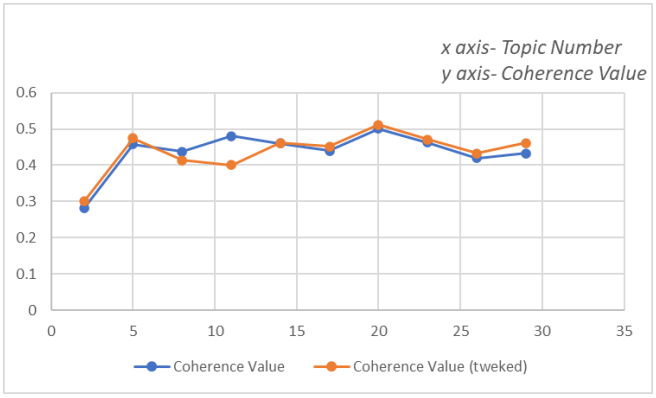


Figure: Coherence Score vs Number of Topics

**For this model the topic coherence score is 0.5120**

The blue color represents the value before changing the alpha and beta values. The other color is after changing the

value. According to the coherence values, when the number of topics is considered as 20, we get the highest coherence values in both cases. So, we will get our best result when the topic number is 20.

## IV. Discussion

We all familiar with traditional machine learning classifiers. Where we split a dataset into training and testing, or we have separate training and testing datasets. After that, we can measure the accuracy of the model as well. But, the LDA topic model does not offer any accurate measurement. We already evaluate the coherence number. The coherence number close to 1, which means the model is more perfect. But the main criteria are whether the topic makes sense. Whether one can identify or classify them simply looking at the topic's word. During this research, we have found some difficulties. And there is some inconsistency in the outcome. In this section, we are going to discuss them.

- Null Value: The Enron dataset offers a great number of emails. And we also choose 20000 emails to analyze. After extracting the email body initially, we lost more than 1000 emails due to the null value in the sender, receiver, or email body field. Then we preprocess the email body. After preprocessing, when we create the corpus, we have seen that some of the email body does not have any value. Because either those emails have junk word or stop words which are not considered. And this lack of word reduces the corpus total word number. Which eventually affect the outcome.

- Word Quantity: From this research, we also learn that LDA does not identify accurately if the number of words in the document is minimal. But it shows good results when the number of words is more significant. So, for the Enron email dataset classification, LDA does not give as many accurate results as it is continuously giving for articles or bioinformatics-related documents. Cause a lot of the emails, consist of an insignificant number of words. And it resulted in the model not performed as expected.

- Word Quality: Enron dataset is an old dataset. The company collapsed in 2001. Later the dataset got published. Although a lot of text-based research has been done based on this dataset, there are a lot of junk words, characters, and typos still available in the dataset. LDA cannot recognize this sort of word. As a result, these unrecognized words also influence the outcome.

- Furthermore, the LDA model does not give the same result, although we do not change any parameter. Because it is a statistical generative model, it predicts the same word in different topic categorizations every time we refresh it. But one thing clear, the LDA model cannot give a good result without a significant number of words and words which can be identified without typos. It may give better results with another email dataset. However, for the Enron dataset, the result is acceptable considering the above factors.

## V. Conclusion

We have evaluated the Enron email dataset to classify email based on the email body contents. We use the LDA topic model to extract human interpretable topics so that we can identify or labialize them in different classes. From our research, we find it easy to implement the LDA model. And we think for the text-based classification LDA topic model will give a better result. At the same time, the LDA model needs a significant amount of data. More quality data will increase the chance of a better result. However, for the email dataset, LDA might work as a quick information retrieval tool that can provide an overview of the email dataset within a short period. But to categorize the dataset, we need to be aware of the word quality and quantity.


## Acknowledgment

I thank Professor Akihiro Fujii, department of Science and Engineering, Hosei University for his guidance and assistance with the understanding of natural language processing and instruction for data implementation. And his continuous support during master thesis. I am very grateful to him for his caring as a supervisor all the time. Then, I am also grateful to Professor Kazuo YANA for giving me the opportunity to be a part of IIST and member of Hosei Family.

Finally I would like to express my gratitude to my parents for their unfailing support my years of study and the process of writing this thesis. I also owe my sincere gratitude to my friends and my classmates who gave me their help and time in listening to me and helping me work out my problems.


## References


[1] David M. Blei , Andrew Y. Ng and Michael I. Jordan, "Latent Dirichlet Allocation," Journal of Machine Learning Research-3 993-1022, 2003.

[2] Ghulam Mujtaba, Liyana Shuib, Ram Gopal Raj, Nahida Majeed and Mohammad Al Ali-Garadi,"Email Classification Research trends : Review and Open Issues" 2013 IEEE Congress on Evolutionary Computation.

[3] Keith Stevens, Philip Kegelmeyer, David Andrzejewski and David Buttler, "Exploring Topic Coherence over many models and many topics," Lawrence Livermore National Lab; Livermore, California, USA.

Griffiths TL, Steyvers M, "Finding Scientific Topics," Proceedings of the National Academy of Sciences of the United States of America, 101, 5228–5235, 2004.