# E-mail communication Analysis using Text Mining

Sachin Malviya

*Scholar, ME (IT)*
*Medicaps institute of Tech.& Mgmt, Indore*

Gogu.Sandeep

*Assistant Professor*
*Computer Science Department,*
*Medicaps Institute of Science and Technology*

## Abstract

*Electronic messages are one of the mostly preferred ways of communication in present days. This paper focuses on analyzing a communication using their textual contents. Analyzing mail communications essential for determining who the parties are between them communication took place and what was the topic on which they were communicating. Online communication (using social networking sites or messengers) analysis is advantageous when we want to detect any criminal communication and also helpful for civilians and military. In this paper we are introducing text mining approaches for analyzing communication via e-messages. By this analysis, we can generate much information about user-behavior, their activities and mostly preferred title of discussions.*

**Keywords:** *text analysis, communication analysis, e-messages mining.*

## 1. Introduction

Presently, internet is a powerful term to connect whole world. Communication among peoples anywhere in the world is possible as they are in front of each other, but here we are analyzing the huge collection of e-messages for collecting information, because the technology brings harm also so that we should not neglect the bad zone of it. Today, communication via electronic messages is a widely used way of connecting to each others.

But we can't identify the communication is of what category, has been done. We need an analyzer which can identify conversation subject and also give information about the behavior of users. User behavior analysis using their text contents is an achievement for us or any social networking site. Illegal activities can also be detected by this analysis. In this paper we are using text mining approach for analyzing communication. Text mining is a field of data mining, also known as knowledge discovery in textual data (KDT). We already know that in electronic mail communication we use textual data; it means that we are sending text messages for communication. Then these text contents can be analyzed using text mining approaches.

## 2. Problems in e-messages communication analysis

We know that mining is a technique, which is used for extracting information from huge data set, but these datasets are arranged in a particular form or follow certain rules. In case of e-mail communication mining we do not have certain data set which follows static rules. E-mail communication, chat or instant messaging does not follow any rules. According to user any sentence could be written in different way. So the structural data sets are impossible to achieve.

The main problem in e-mail communication mining is its language. When two or more users communicate via instant messaging system then their language of communication is probably different from conventional text. Because in this type of communication, users use short texting format, acronyms, wrong spelled words and usually repeating last letter more times like "hiiiiii" instead of "hi". Such kinds of terms used in electronic text messaging are describes below.

I) **Short forms of words**: we use short forms of words in electronic text communication, or sometimes only initial letters are typed instead of typing whole word like

See you – CU
Are you okay – r u ok?
How are you – hru?
Fine - F9
Great - G8
Example- e.g.
The- d
Girlfriend/boyfriend- GF/BF

II) **Wrongly spelled words:** Usually we repeat last letters many times in text email communication, hiiii, byeeee, noooo, sooo and also many wrongly spelled words like nyt instead of night, are typed by us in text communication, it's a habit.

III) **Sign Language**: In text communication via online medium we often use smilies for expressing our feelings. These signs are kept an important existence in textual communication.

☺, ☺ for happiness or smiles
☹, ☹ for unhappiness or sadness
**:**p, **:**D for showing tongue

IV) **Multiple topics**: It is not necessary that only a single topic is discussed in the communication, so many topics are randomly changed and discussed in the communication, so that this is another problem to find a specific topic from whole conversation.

These are the some problems we have to face in analyzing a textual mail communication.

## 3. Methodology

As we previously discussed instant messaging or chat language is quite different from conventional communication text. Here we are using text mining techniques for analyzing text communication over internet. For understanding the communication title or what discussion took place in that communication, we have to consider different mining techniques.

We know that in data mining or knowledge discovery in database firstly we apply preprocessing techniques for cleaning the data set and for finding only required piece of data. Same thing here, we are using text preprocessing techniques first for extracting essential text for classification and mining.

Text preprocessing is the technique of cleaning textual data from text communication. Cleaning means, it removes all the spaces, articles, special characters from the textual contents. So that size of all the textual data set is decreased by applying this technique. Specially techniques used in this textual content preprocessing are:

(a)Stemming, (b) Lemmatization and (c) filtering

a) Stemming is the technique of extracting the original form of words. For ex

***Actual word***

Working, Works, Worked - Work
Beginning, Begins - Begin

b) Lemmatization is close to the stemming but in this process may also involve complex task like knowledge of context and parts of speech. It is more difficult than stemming process. For example words like disputing, disputed, disputes would be considered as disput in the stemming process but their normalized form is dispute. That only can be obtained by lemmatization process.

c) Filtering is the technique of removing complexities of textual sentences. Complexities means words which can be removed from sentence without any information loss, like articles (a, an, the), prepositions.

I AM VERY HAPPY.

If we have only two words I and HAPPY, then we can understand the feeling of user, that he is happy now.

With text preprocessing we apply different other text mining approaches for mining the textual contents writing by users. These approaches are:

i) Divide Collection of messages into short length pieces
ii) Extraction of text features
iii) Selection of text features
iv) Supervised grouping of texts
v) Unsupervised grouping of texts
vi) Creating summary of extracted texts

### I. Divide collection of messages into short length pieces

We have large amount of message data sets. Those data sets are collection of several different sessions of communication. We need to shorten those contents so that further text analysis processes can be applied efficiently. These short pieces can be divided according to time duration, for example we take 10 minute text contents as a small session or we can also consider number of words for deciding session length. After that we apply other techniques for extraction of features classifications.

## II .Extraction of Text Features

A session may contain numbers, icons, e-mail addresses, URL's, special characters with textual contents. In this step we extract different features and categorize them.

## III. Selection of text features

Feature Selection in textual data is performed by indicative terms that are stored in the indicative term dictionary. It is used for classification purposes to select appropriate features. That is based on our examination that chat conversations on a particular subject usually contain a set of words, known as *indicative terms* (or *topic keywords*) that characterize that particular topic. This set of indicative terms is considered to be highly representative for all conversations on the same topic. Therefore, indicative terms can be treated as a unique collection Of features characterizing the chat contents belonging to a particular topic Indicative terms are not limited to single word, it might be phrases as well. With indicative terms predefined as features for selection, it can also reduce the dimensionality of input features to the classifiers. Feature Selection process consists of two steps Tokenization and Indicative Terms Identification. Tokenization simply breaks the chat session content into a list of single words. Lower cases are preferred for preprocessing; firstly each term is converted into lower cases for processing. Then indicative Terms Identification selects a set of terms from the tokens for each topic category based on the Indicative Terms Dictionary. The selected indicative terms will then be incorporated into a feature vector which will be used as the input to the topic classifiers. The weight of each indicative term in the feature vector will be 0 or 1 depending on the appearance of the corresponding term in the chat session. As only the indicative terms will be used for the categorization process, it is of the utmost importance to choose the most representative set of indicative terms for each topic category. Topic category list should not be very small or very long. If that topic category list is small then performance will be affected and if it is long then different overheads may be occurred. Using training data set we can understand the behavior of analytical method and define different categories as topics.

## IV. Grouping of Text

Three sets of classifiers have been built based on three different classification techniques, Naive Bayes (NB), Associative Classification (AC), and Support Vector Machine (SVM).

## 4. Future work

In the next work we can understand the output of text mining using different techniques. Some statistical data analysis methods can be used for our task, these methods will understand the output of the mining tool. By this mining tool we can analyze behavior of user who is communicating.

## 5. Conclusion

On the basis of e-message analysis results, we have proposed an indicative term categorization approach for chat topic detection which incorporated different techniques such as sessionalization of chat messages and the extraction of features from icon text and URLs for pre-processing step. Some different classification techniques such as Naive Bayes algorithm, Associative Classification, and Support Vector Machines are employed as classifiers for observing title of the communication, proposed approach has been evaluated based on precision, F-measure and accuracy from the other e- messages data set collected from the Web. Moreover, the proposed approach is shown superior than the document frequency based approach and is highly computational efficient as it is able to achieve, relatively high and stable performance with just limited number of features, which makes it suitable for online monitoring.

## 6. References

[1] Jie Tang, Hang li and Zhaohui Tang "Email Data Cleaning."

[2] Dr. Anadakumar. K and Ms. Padmavathy. V "A Survey on Preprocessing in Text Mining *"International Journal of Advanced Research in Computer Science, vol4 2013*

[3] Michal Tomana, Roman Tesara and Karel Jezeka "Influence of Word Normalization on Text Classification"

[4] Chee Wee Leong and Rada Mihalcea and Samer Hassan "Text Mining for Automatic Image Tagging."

[5] Sanjay Tanwani and Neha Rahatekar "Automated Personal Email Organizer with Information Management and Text Mining Application"

[6] Tianhao Wu, Faisal M. Khan, Todd A. Fisher, Lori A. Shuler and William M. Pottenger "Error-Driven Boolean-Logic-Rule-Based Learning for Mining Chat-room Conversations"