IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Classification, Summarization and Event Extraction through Emails

Shubham Kulkarni[1], Amartya Pandit[2], Mayur Waghmode[3], Arnav Desai[4], Prof. Mrs. S. Chitale[5]

B.E.(Computer Engineering) Student, Pune Vidyarthi Griha's College of Engineering and Technology, Pune-411009, Maharashtra, India [1,2,3,4]

Professor, Department of Computer Engineering, Pune Vidyarthi Griha's College of Engineering and Technology, Pune-411009, Maharashtra, India[5]

Abstract : **In this digital era, a lot of information is conveyed through emails. The email service provider applications show the users their email data but do not process it further for more convenience of the users. Automatic event generation and detailed email classification are absent in the conventional email applications. In this paper, an approach for the classification of mails, summarization and automatic event extraction through emails is discussed. The emails are classified into multiple classes using Machine Learning and not only in spam and ham. Rule-based extraction is used for summarization. The date, time, subject and venue of the event present is extracted through emails using text mining. The application is evaluated over many emails and is providing satisfactory results**.

Keywords : NLP, SVM, Multiclass Classification, Extractive Summarization, Event Extraction

I. INTRODUCTION

An email user receives a minimum of 15 emails as an average per day. These mails belong to many categories like the mails from his/her organization, offers from various services, promotions, job descriptions etc. Some of them might be important to the user and some of them might not be. Also, sometimes the user may forget to read some emails and important events like a meeting or bill payments are missed. Hence, it is of utmost importance to organize the user's inbox into multiple categories so that he can easily ignore unimportant emails. He will not have to waste much of his time reading lengthy emails if a summary of those emails is shown to him. He will not miss the important events if a reminder is given to him previously. Hence, we found that an Android application will be quite helpful to solve the described problem since a considerable number of email users use an Android smartphone.

The main focus of this paper is classification, summarization and event generation through emails. Natural Language Processing (NLP) is a popular field in computer science nowadays. It is used in the proposed approach for the summarization and event generation process. Every email contains some important keywords which are responsible for its category. These keywords can be found out using the machine learning approach for the classification. Extractive rule-based summary generation approach is used for the summarization. Every event in the email contains some typical information like the event date and time. This data can be extracted using NLP techniques. The user can get the important information without reading the whole email using the proposed approach.

II. RELATED WORK

This paper discusses an abstractive dictionary-based approach for the summary generation. A user-specific dictionary is developed by the system which contains the most preferred words by a user. The mail subject and body is passed to this dictionary. It determines what the mail is actually about and constructs an abstractive summary. The system also makes the use of common words in the mail body to determine the mail's class [1]. In this paper, an email classification approach is developed using Support Vector Machines and Naive Bayes classification. The mails are classified into spam and ham with an accuracy of 93%. Rough Set theory is used to reduce the number of attributes for classification. 58 attributes are used for Naive Bayes and 9 attributes are used for SVM classification [2]. To retrieve information about multiple emails into a structured format, this paper has discussed an extractive summarization approach which is not rule-based. Sentiment analysis is performed on the mail body to decide the emotion behind the mail and the related sentences from the body are extracted as the summary [3]. A statistical extractive rule-based approach for summary generation is discussed in this paper. Each sentence in the mail is assigned with a numerical value based on certain features like Title Feature, Numerical Data Feature, Proper Noun Feature etc. Each sentence is passed through all these rules and a score is assigned to the sentence. Top 20% sentences are assigned as the summary in the same order as they appear in the original mail [4]. Naive Bayes and Support Vector Machine classification are efficient machine learning algorithms for text classification. This paper compares their performance over multiclass document classification to find out that Support Vector Machines yield more accuracy than the Naive Bayes classification [5]. In this paper, an Android application using Kivy framework has been proposed to extract event information from the email body. Date, time, venue and subject of the email is extracted using ANNIE NLP framework and stored in the phone's calendar [6]. This paper introduces some new rules for the extractive summarization approach. The algorithm asks the user for the number of sentences in the summary and presents that number of sentences with the highest scores [7]. This paper proposes an extractive summary generation approach using sentence scoring method. After the preprocessing consisting of stopword removal, tokenization and stemming, each sentence is assigned with a score after passing through various rule features viz. Title, Term-Weight, Sentence Position, Date-Time, Proper Nouns, Numerical Data and Topical Words [8].

III. PROPOSED APPROACH

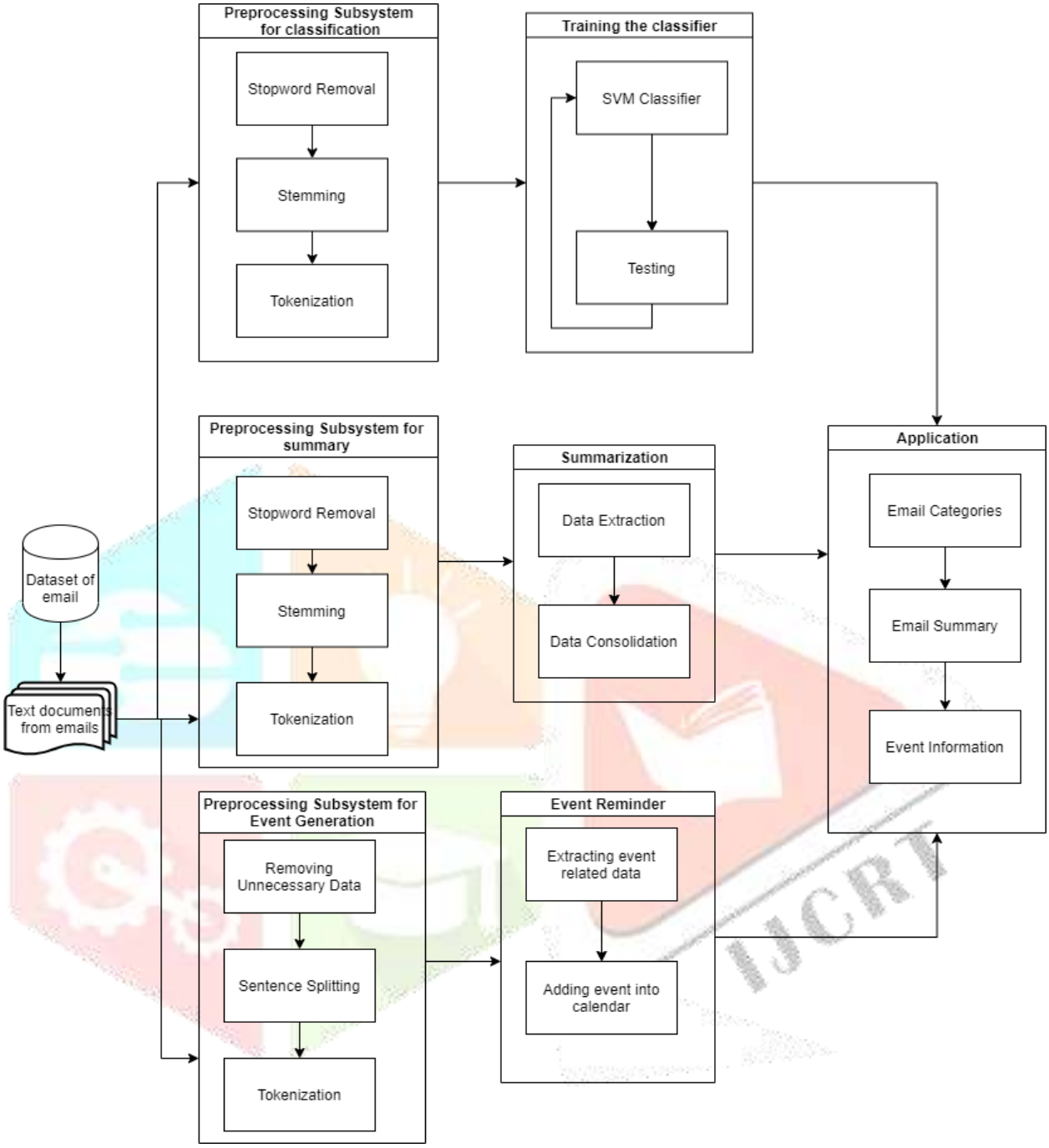

Fig. 1 System Architecture

The proposed system approach has three main modules: Classification, Summarization and Event Generation as shown in the System Architecture. An Android smartphone is used as the user interface.

Classification

For categorization, we formulate the task as a classification problem to classify the mails into five major classes and their subclasses as follows using the Machine Learning approach. The classifier is trained on a dataset containing 800 emails.

- Primary Mails
- Educational (Courses, Exams and Competitions, Admission)
- Offers (Clothing, Travel and Hotels, Food)
- Employment (Jobs, Internships)
- Finance (Transactions and Statements, Investment).

Term Frequency – Inverse Document Frequency (TF-IDF) transformer has been used for the classification. Support Vector Machines classification approach is used due to its proven performance in the document multiclass classification[5]. We have used the dataset of emails collected by us to train the classifier.

Proposed Algorithm for Classification :

- Start
- Accept the mail body
- Remove stopwords
- Perform stemming and tokenization.
- Provide pre-processed mail body to CountVectorizer to assign count to word according to its repetitivity
- Feed the count values of mail body to TF-IDF transformer to convert the count to create biases to count to mail body compared to count of words in other mail bodies
- Provide the array of biases to SVM Classifier which divides theses different categories in form of hyper-planes
- Update the category of mail.
- Stop

Summarization

For summarization, an extractive approach with sentence scoring method is used. OpenNLP, which is a machine learning based Natural Language Processing toolkit, is used for this purpose. Named Entity Recognition is carried out using OpenNLP. Each sentence is passed through certain features (rules) and if a sentence matches a certain rule, some score is given to that sentence. All the features are calculated after the stopword removal and stemming. The features along with their scores are as follows:

- *Subject Similarity Feature*
  This feature calculates the similarity of a sentence in the body with the subject of the email.
  It is calculated as: (No. of subject words in a sentence / No. of words in the subject).

- *Numerical Data Feature*
  This feature takes into consideration the numerical data in a sentence. Numbers written in words (e.g. ‘one’, ‘two’ etc.) are also considered as numerical data.
  It is calculated as: (No. of numerical data in a sentence / 10)

- *Link Feature*
  This feature takes into consideration the presence of any web address in a sentence.
  If a link is present in a sentence, its link feature is set to 0.5, else its value is set to 0.

- *DateTime Feature*
  Presence of date and time in a sentence makes it important for the summarization.
  If only date or only time is present in a sentence, then this feature is set to 0.5. If both of them are present then this feature is set to 1. Else, this feature value is set to 0.

- *Topical Words Feature*
  The words which are repeated at a considerable extent throughout the mail are topical words. Each word in the mail is assigned the value equal to its total appearances in the mail body.
  This feature value is calculated as : (sum of values of all words in the sentence) / n
  where, n = total number of sentences in the body

- *Proper Nouns Feature*
  The sentences containing a proper noun are important sentences in the email.
  Hence, for each proper noun appearing in a sentence, 0.5 is added to this feature of the sentence.

Proposed Algorithm for Summarization:

- Start
- Accept Email subject and body.
- Split the body into sentences and then into tokens.
- Remove stopwords from subject and body.
- Calculate each feature value for each sentence.
- Arrange the sentences in descending order of their scores.
- Select 40% of the total number of sentences from the top.
- Arrange them in order of their appearance in the original body.
- Assign these sentences to the summary variable.
- Return summary
- Stop

Event Extraction

Spacy, a Natural Language Processing library is used for event extraction instead of the conventional NLTK toolkit since the former is more efficient and accurate than the latter. The email body is split into sentences. Event extraction is carried out only on those sentences which contain a date in them. The date and time and venue are searched using Spacy as well as manual programming approach for more accuracy. The proposed algorithm is able to extract multiple events from one email. The event could be anything like a meeting, a deadline, an exam etc.

Proposed Algorithm for Event Extraction:

- Start
- Accept email body and subject.
- Remove unnecessary punctuations.
- Split the email body into sentences.
- For each sentence, search for the date using Spacy as well as manual approach (regular expressions)
- If a date is found, append the date to the dates list.
- Search for the time in the same sentence and in the next sentence.
- Append the time found to the times list.
- Append the noun phrase coming after the words ‘at, in’ in the same sentence to the venues list.
- Append the subject of the mail with the first noun phrase and verb in the same sentence to the subjects list.
- If time and venue is not found, append blank string to the respective list.
- Search for the word “Venue” in all sentences.
- If found, replace all the contents of the venues list with the remaining sentence.

- Convert all contents of the dates list to a specific date format using date parser.
- Return events
- Stop

IV. EXPERIMENTAL RESULTS

Classification

Classification module is trained using a dataset of emails collected by us and with the help of CountVectorizer and TF-IDF Transformer to convert the words in the mail body into numerical values that can be understood by the machine. Accuracy achieved by the SVM Model is about 88%. Further these categories are sub-categorised using different models trained for the specific category recognition. Providing users with in-depth categorization, which no other email clients do not provide. Categories provided by the module are as follows. Subcategories are mentioned in the brackets.

1. Education (Courses, Exams - Competitions, Admissions)
2. Finance (Bank Statements, Investment)
3. Jobs (Jobs, Internships)
4. Offers (Clothing, Travel - Hotels, Food)
5. Primary

Category	Precision	Recall	F1-Score	Support
Education	0.89	0.78	0.83	128
Finance	0.92	0.98	0.95	59
Jobs	0.79	0.90	0.84	115
Offers	0.95	0.92	0.94	158
Accuracy			0.89	460
Macro Accuracy	0.89	0.90	0.89	460
Weighted Accuracy	0.89	0.89	0.89	460
Accuracy Score				0.886956521739130

Table 1. Classification Accuracy Report

Summarization Module

We have compared the extractive summary generated by the proposed approach to the extractive summary generated by experts for 110 long emails (words > 100) from the collected dataset. It was found out that the algorithmically generated summary was 83% similar to the summary generated by experts. The summary generated is informative and meaningful as per expected.

Event Generation Module

The event extraction model is expected to extract subject, date, time and location of the event. Some of the events identified by the proposed algorithm for the given input are:

- Input : The meeting is scheduled on Tuesday, 11th Feb 2020 at the library's reading hall.
- Output :

Subject	Date	Time	Venue
The meeting schedule	11/02/2020	Null	library's reading hall

- Input : CoCubes is conducting the test on 23/01/2020 at 9:00 AM.
- Output :

Subject	Date	Time	Venue
CoCubes conduct	23/01/2020	9:00 AM	Null

- Input : I will distribute the hall tickets tomorrow during lunch hour. Please be present.
- Output :

Subject	Date	Time	Venue
I will	23/04/2020	lunch hour	Null

The algorithm has been tested with 112 emails out of which 76 emails were having events and remaining were not. 66 events from 76 emails were recognized correctly. From the 36 emails not consisting of any event, no event was detected by the algorithm as expected.The algorithm recognized cases of relative time durations and date formats.

TYPE OF EMAIL	ACCURACY
Mail with complete event information	88%(44 out of 50)
Mail with incomplete event information	84%(22 out of 26)
Mail with no event data	100%(36 out of 36)

Table 2 Event Generation Accuracy

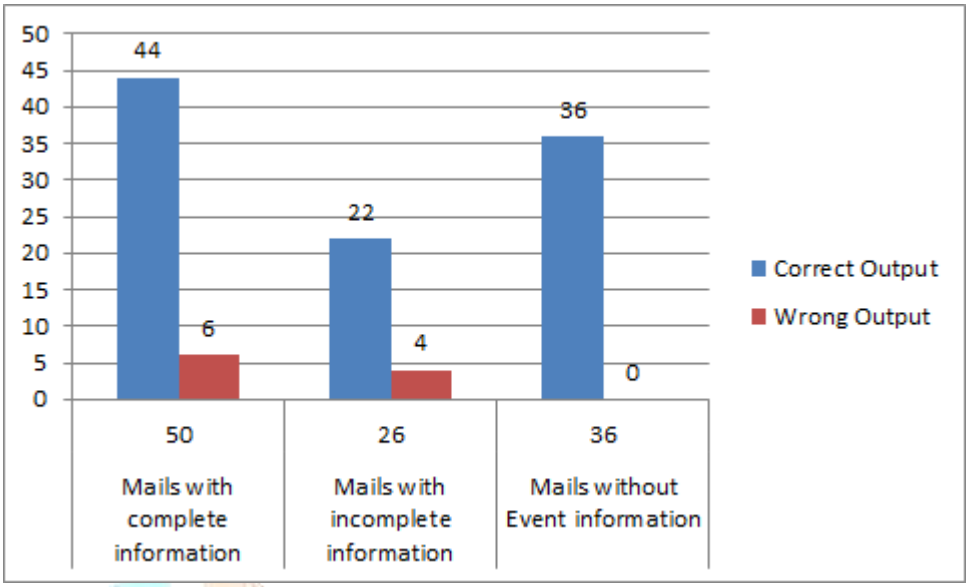

Fig. 2. Event Generation Accuracy Graph

V. CONCLUSION and FUTURE WORK

A large number of emails pile up in the user's inbox. There is a need to organize the inbox and extract the useful information from the emails for the ease of the email users. Our proposed approach helps for the detailed multiclass classification of the emails into predefined categories. Informative summary of the emails is also generated and information related to an event is extracted. Multiple events present in an email are also extracted.

The summarization algorithm uses a statistical extractive approach with predefined features. The performance of summarization could be increased by adding some more features and combining an abstractive approach to the current one. Machine learning can be used to extract the venue and subject of an event instead of the static approach used in the event extraction algorithm. For the classification module, more categories and sub-categories could be included and accuracy of its classification could also be improved with help of a large and diverse dataset of emails.

VI. REFERENCES

1. Taiwo Ayodele, Shikun Zhou, Rinat Khusainov. "Email Grouping and Summarization: An Unsupervised Learning Technique" : *In 2009 World Congress on Computer Science and Information Engineering*
2. Zhiqing Zhu. "An Email Classification Model Based on Rough Set and Support Vector Machine" : *In Fifth International Conference on Fuzzy Systems and Knowledge Discovery*
3. Ashraf Q. Mahlawi and Sreela Sasi. "Structured Data Extraction from Emails" : *In 2017 International Conference on Networks & Advances in Computational Technologies (NetACT) (20-22 July 2017)*
4. Siya Sadashiv Naik, Manisha Naik Gaonkar. "Extractive Text Summarization By Feature-based Sentence Extraction Using Rule-based Concept": *In 2017 2nd IEEE International Conference On Recent Trends in Electronics Information & Communication Technology (RTEICT), May 19-20, 2017, India*
5. Zun Hlaing Moe, Thida San, Mie Mie Khin, Hlaing May Tin. "Comparison of Naive Bayes and Support Vector Machines on Document Classification" : *In 2018 IEEE 7th Global Conference on Consumer Electronics (GCCE 2018)*

6. Aneesh G. Nath, Krishnanth V, Kevin Biju Mathew, Pranav T S, Sarath Gopi. "NLP Based Event Extraction from Text Messages" : *In International Conference on Future Technology in Engineering – ICFTE 2016*
7. T. Sri Rama Raju, Bhargav Allarpu. "Text Summarization using Sentence Scoring Method": *In International Research Journal of Engineering and Technology (IRJET) (Volume: 04 Issue: 04 | Apr -2017)*
8. Mithak I. Hashem. "Improvement of Email Summarization Using Statistical Based Method" : *In International Journal of Computer Science and Mobile Computing, Vol.3 Issue.2, February- 2014, pg. 382-388*