IEC
Institute of Electronics
and Computer

Journal of Artificial Intelligence and Systems, 2021, 3, 93-114
https://iecscience.org/journals/AIS
ISSN Online: 2642-2859

Feature Extraction aligned Email Classification based on Imperative Sentence Selection through Deep Learning

Nashit Ali[1], Anum Fatima[1], Hureeza Shahzadi[2], Aman Ullah[3], Kemal Polat [4, *]

[1] Department of Computer Science, COMSATS University Islamabad, Vehari Campus, Vehari 61100, Pakistan
[2] Department of Computational Science & Engineering, National University of Sciences and Technology, Islamabad.
[3] School of Computer Science and Engineering, Central South University, Changsha, 410083, China
[4] Department of Electrical and Electronics Engineering, Bolu Abant Izzet Baysal University, Bolu 14280, Turkey
*Corresponding Author: Email: kpolat@ibu.edu.tr (Kemal Polat)

How to cite this paper: Nashit Ali, Anum Fatima, Hureeza Shahzadi, Aman Ullah, Kemal Polat (2021). Feature Extraction aligned Email Classification based on Imperative Sentence Selection through Deep Learning. Journal of Artificial Intelligence and Systems, 3, 93–114. https://doi.org/10.33969/AIS.2021.31007.

Received: July 27, 2021
Accepted: August 13, 2021
Published: August 16, 2021

Abstract

Most commonly used channel for communication among peoples is emails. In this era where everyone is so busy in their routine and work, it is very difficult to check all email when one receives huge amount of emails. Previous research has done work on email categorization in which they have mostly done spam filtration. The problem with spam filtration is that sometimes person mistakenly mark an important email received from high authority as spam and according to previous research, this email will be filtered as spam that can cause a great threat for job of an employee. In this research, we are introducing a methodology which classifies email text into three categories i.e. order, request and general on basis of imperative sentences. This research use Word2Wec for words conversion into vector and use two approaches of deep learning i.e. Convolutional neural network and Recurrent neural network for email classification. We conduct experiment on Dataset collected from Personal Gmail account and Enron which consists of 1000 emails. The experiment result show that RNN gives better accuracy than CNN. We also compare our methods with previously used method Fuzzy ANN results and Our proposed methods CNN and RNN gives better results than Fuzzy ANN. This research has also included different experimental result in which CNN and RNN applied on different ratios of training and testing dataset. These experiment show that increasing in the ratio of training dataset results in increasing accuracy of algorithm.

Keywords

Convolutional neural network, Email Categorization, Imperative Sentences, Recurrent neural network, Spam filtration.

1. Introduction

Emails is considered as easiest and fastest means of communication among people. the usage of emails has been increased since many decades. People feel comfort to sharing information, personal data, professional data or orders through emails because it is considered as cheapest and fastest means of communication among people. According to a report named Email Statistical Report 2016-2020 [1] by Radicati Group Inc., it is presented the fact that use of email for different purposes continue to grow all over world. According to this report, "there were 2.6 billion active email users in 2016 and there will be 3.0 billion email users in 2020 according to their expectation. The expected number of business and consumer email sent each day will increase with an annual rate of 4.6%". in 1996, Sinder and Whittaker Introduced the term email_overload [2]. Email_overload refers to using email for those activities it was not designed for such as meeting arrangements, management of useful information regarding work and contacts. Email overload is also defined as one incapability to check and process all received emails. This leads to anxiety and became inefficient as large amount of time in consumed in checking and managing emails. An automated algorithm or technique that arrange emails in different categories will be solution of email_overload which reduce anxiety and increase efficiency. The solution is known as email categorization. As trend of using technology continues to grow worldwide, peoples are completely depending upon technology. Emails gives a lot of comfort and benefits to people but it is also having a bad side as it is used for wrong purposing such as viruses, spam emails. These emails can be used for fraud in banking and advertisement or viruses. These emails received from unknown sender. An active filtering method required to avoid unwanted emails. Arrangement of email is known as filtering which is also known as email categorization.

With the increasing trend of using internet, people feels more comfort to publish their information on internet and share data with others through internet instead of manually sharing. Usage of technology without any proper way or technique is just waste of time. Usage of technology need a proper technique so that it gives maximum benefits to users. Same situation can be considered in emails as emails surely provide comfort to people but still there is lot of things that's need to be done by themselves. If the number of emails is small, then it is ok to do arrangement, checking tasks manually but if the email account contains large number of emails then these tasks become inefficient and tiring. Therefore, there is need of automated emails categorization so that it saves users time and reduce the possibility of losing any important email due to receiving of large number of emails in same time. User mostly use emails to save information of their as in case if they need it they can find it in emails. But if the account contains thousands number of emails, the required email got lost and user become tired on finding it whenever he need it. So there is need to divide emails in to folders or categories so that it can be retrieved whenever needed without consuming time. For this purpose, an algorithm is needed that automatically categorize email, this categorization can be done on basis of some features such as body of email, subject of emails, relationship between user and sender. For example, electricity and other bills should be in one folder etc.

Converting input data into set of feature is known as feature extraction. Emails content is converted into feature then sentiment analysis is done on it. Emails are categorized on these features. As people use emails to save their information and to retrieve it whenever they need. The increasing usage of emails also increasing difficulty of user as their they receive hundreds of emails daily so the required email got lost. To avoid this problem, email should be categorized in folders. The categorizations should be done on some features, for example a single user receive hundreds of emails daily including email from his boss but due to lack of time he is not able to oversee and process all emails which cause a risk to his job. To avoid these types of problem, email should be categorized on basis of features. Features can be importance of content, relationship between user and sender, topic of email etc. in this research we are categorizing emails on features such as order, request and information. English language is full of ambiguity for example “Please is used for request as well as formal order” such as an employee receive an email from relative “Please save this work” and also receive an email from boss “Please fill Performa and save it in documentation”. In both emails “please” is used but in first it is considered as request but in second it is an order. feature based sentiment analysis is used to reduce ambiguity and categorize emails on basis of separate folder so that important emails will not lost which cause in loss for user later.

The main problem in this research is generating feature representation that retains significant features in a lower dimensional feature space for prioritizing emails. The Significance of this study to do email categorization on basis of feature based SA. The major contribution of research is categorization of emails according to the user prioritization features such as order, request and information. Many researcher works on email categorization but these features remain untouched. This research could be a new way of thinking to researchers.

This research is done in order to categorize emails in to order, request and general category. In [3] previous work on email categorization has been discussed and future trends also has been discussed in which this article conclude that there is still need of improvement in email categorization so need to focus work which is not done yet, this article gives an idea of updating feature on which email can be categorized. In article [4] email classification has been done on dataset email account of an e-commerce website. This article classifies emails on five features happy, worst, satisfy, disappointed and okay. this scale has been made for showing customer satisfaction towards their product or service. This article used ANN for email classification. After studying these articles, I came with proposed article which categorizes the email on different sentiment features that include order, request and information using Deep learning algorithms. This topic aims to work on attitude of sender in email It judges email if someone is making request, giving order or giving important information.

Following are our research objectives

- Compare Deep learning techniques
- Testing the accuracy to find out better algorithm
- To categorize emails on basis of feature based sentiment analysis.

In this research we have considered these research question:

Q: Does previous categorization prioritize your email by proposed feature of this research?

Motivation: This will explore the need of proposed features in this research as the previous classification remove the spam or junk emails but sometimes email from high authority can be placed mistakenly in spam hence the previous researches also removed important email of yours which may cause in your loss.

Q: Does the proposed approach useful to user?

Motivation: This will help researcher to do further research on this aspect of email categorization

Our proposed research is limited to feature space in which we are using only three feature i.e. order, request and information to categorize emails. Some accent is still under process for example in email text, “Do it” is interpreted as order but in real if “Do it” is pronounced in low key then it does not interpret as order so these kind of accent are still under process. As we know English language is an ambiguous language in which one word has multiple semantics in different context. This research reduces the ambiguity of emails as understanding the context of emails and categorize it. This research will help people as they need less time to find emails which can cause problem for them such as an email from high authority officer.

Our methodology to collect literature review includes three steps which are:

- Searching
- Criteria for exclusion.
- Criteria for inclusion

Searching

We searched necessary papers which have frequent citations from many different online sources like IEEE, Springer, ACM, Google Scholar, Science Direct. We used sentiment analysis, feature extraction in sentiment analysis, email categorization, feature based email categorization etc. as keywords for searching by which we get maximum papers which we required for our research.

Inclusion Criteria

Only Research articles and thesis which were published from 2014 to 2018 are counted in this research. Those publications which has frequent citation but published before 2014 are also used in this research, as they deliver basics of knowledge and clear understanding to problems under about required issues.

Exclusion Criteria

Any kind of unpublished material, Abstracts, reports and thesis are not used in this research for literature review. Those publications which are written in any language other than English or which do not English translation are also not used.

2. LITERATURE REVIEW

Now days user consider email as one of important way of personal and business link of people. As the volume of emails increases, it is important to categorize emails to save time, to avoid spam emails and for many others reasons. However, email categorization has been become an appealing topic for research. Ghulam Mujtaba and co-authors presented a research article which broadly reviewed articles published in 2006-2016 on email classification [3]. They exploit their analysis in five phases: data sets that are used in e-mail classification area, application area, e-mail classification methods, feature space used in every field and the use of performance measures. Research challenges, issues and gaps were also presented in this article for direction of future researcher. According to this article, email classification has been used in fifteen application areas which includes spam email categorization, multi-folder email categorization, phishing emails etc. According to article, the most used feature used in email classification are body content, header part, URL and JavaScript of email etc. [3]

Managing emails is become a considerable issue now. Many methods are used to classify emails such as statistical Bayesian, Naïves Bayes algorithm etc. These algorithms use difficult artificial intelligent techniques. These algorithms have drawbacks such as low accuracy, less efficiency and not handling sarcasm as elaborated in article written by Akash Kumar Singh, Akshay Nair, Krishnakant Mahto, Kundan Gadgil and Prashant G. Ahire [4]. This paper proposed an approach towards building a model using NLP, Fuzzy Artificial neural network(ANN) and machine learning techniques for classification of emails using pre-defined protocols. In this article, emails from Gmail is used as data set in this system, Fuzzy ANN algorithm is used because according to author there is less work done in email categorization is with help of Fuzzy ANN. This algorithm converts the extracted feature into numerical score then on basis of these score, values are arranged according to fuzzy ranges then conditions apply on it which categorize emails. This article concludes that the proposed system produces better result with high values. [4]

People trust emails as one of the most secure communication medium for transferring data and information. The volume of unwanted data grows rapidly with increase of population so different filtering method developed by researcher that filter these unwanted data or massages. There exist many spam detecting techniques that includes Heuristic processes, Knowledge-based technique, Clustering techniques, Learning-based technique etc. Hanif Bhuiyan, and co-authors presented a study [5] of different techniques used for spam email detection i.e. “Naïve Bayes, SVM, K-Nearest Neighbor, Bayes Additive Regression, KNN Tree”. This article compares and evaluate these systems and concludes that most of filtration is done through Naïve Bayes and SVM. Each technique has effective outcome but have loop holes for researcher to increase their performance. [5]

Classification of anything whether e-documents or emails require NLP techniques as well as ML techniques So R Manikandan and Dr. R Sivakumar presented review on ML algorithms for text classification [6]. This article presented different methods such as NB, SVM, DT, Decision rules Classification, Rocchio’s algorithm, K-Nearest Neighbor, Fuzzy Correlation and Genetic Algorithm along with their advantage, disadvantage and applications. Among algorithm discussed in this article, Naïve Bayes, K-NN and SVM is concluded as most appropriate algorithms for classification where other algorithms can give efficient result in combination with others. There is need for improvement in these algorithm so that they can give optimal results. [6]

Since last decades, Sentiment analysis has been an appealing topic for researcher. Research has been done on social media blogs and other online documents but SA of email has not been as studied as it should be because it itself an important topic. Authors proposed a hybrid framework for SA using TF-IDF for purpose of extracting feature, on email dataset and then a k-means (hybrid) with SVM classifier for classification that gives better output as compared to other combination of algorithms. [7] For Conducting Experiments, firstly they compare 3 methods of each feature extraction and clustering and 5 algorithms of classification to justify why they chose proposed model. Feature selection methods includes “Bag of Words, term presence and frequency-inverse document frequency”. Sentiment clustering method includes “polarity labeling. sentiment classification methods include SVM, NB, LR and DT(J-48)”. Experiment result concluded that term presence and frequency-inverse document frequency model performs better that Bag of Words in case of extracting feature, k-means outperforms other for clustering and SVM reliably perform better for classification. [7]

Technology has given great comfort to world but also technology can be used for wrong intentions. For example, spam emails are example of wrong usage of emails that cause discomfort to peoples as they keep receiving spam emails without their concern or wish. That’s the reason researcher keeps on finding method that classify spam emails to save users. Esha Bansal and Pradeep Kumar Bhatia present comparative analysis of pre-existing classification techniques on basis of their performance [8]. This article concluded that there is need to use feature selection technique to reduce training time and ensemble based techniques Boosting to improve the accuracy. So, feature selection algorithm and ensemble based techniques should be combined for better efficiency. [8]

As previous literature discussed that performance of classifier can be improved. Ensemble classification techniques in machine learning algorithms are used for improvement in performance of classifiers. P. Visalakshi and co-authors proposed a system named as “an Ensemble Classifier for Email Spam Classification in Hadoop Environment” [9]. Gradient boost ensemble technique is used with DT and NB algorithm. These Ensemble Algorithms combines set of weak learners and results in improved predicting accuracy. The proposed model involves two phases i.e. train & test. This article presents the performance measures i.e. Naïve Bayes gives 80% accuracy and 80% precision but gradient boosting algorithm gives 94% accuracy and 92% precision so concluded that the proposed system improves the performance of classifiers [9]

Spam emails has various disadvantages such as it reduces productivity, takes extra space in mailboxes, extra time, spread viruses, and contains data that can destroy Internet clients and servers. Shradhanjali and Prof. Toran Verma proposed approach using SVM and feature-extraction for detecting email if it is spam or not which gives accuracy of 98% [10]. The methodology consists of pre-processing, extracting feature, SVM training, test classifier, test email. Pre-processing step removes stop words urls, numbers and special character then do word stemming. Feature extraction is extraction of meaningful words from the text which later mapped from vocabulary list. Emails dataset is used to train classifier after that it is ready to classify emails. Then the testing phase test classifier with sample data and in the final stage emails are given as input and classifier gives output in binary numbers as 0 means email is not spam and 1 means email is spam. [10]

In most classification technique, smaller context of sentence has been concerned so Xingyi, Johann Petrak and Angus Roberts proposed a new model which is named as Context_LSTM_CNN model [11]. word embedding is used to convert sentence into words then on these words bi-directional LSTM is applied in this model, CNN is now applied on result of previous step. After that the FOFE is applied to both context of word embedding which comes in form of final result. [11]

Email inbox consists of important messages but also contains unwanted emails which consume time space and bandwidth. To avoid these unwanted emails, there were many techniques available to classify emails. Priti Kulkarni and Dr. Haridas Acharya compares various email classification techniques that use email header fields as feature to classify. [12].   This article concludes that with header features decision tree and K-NN algorithm gives best results among all and bagging gives lowest results. Bayes net outperforms all other classifier. Performance of classifier is evaluated using “accuracy, true positive (TP), false positive (FP), precision, and recall". This article concludes that DT J48 outperforms all classifier but random forest algorithm and bagging perform poor among all the classifiers. [12].

E-mails is frequently used in organizations and for business purpose and it is considered as ubiquitous and secure mode of communication. Researcher find out many techniques to classify spam emails in to spam folder. Message in emails is written in natural language so one line of text can contain more one than questions and one word can contain more than one meaning according to its context. Automatic replies to emails is useful in organization and institute where they have to forward email to expert or reply to their customers. Classification method can be used to classify in to user-defined multi-folders as proposed by [13]. They used Naïve Bayes and SVM method and also discuss effect of PoS tagging and lemmatization on Precision of Classifier which results in 81.7% precision with Naïve Bayes and 85.3% with SVM. This article concludes that Lemmatization reduces precision and recall but PoS improves overall result. [13]

Many methods are available to filter spam emails. Artificial neural network is considered as powerful method to classify emails as it has capability to results in better accuracy even with the huge amount of dataset. Mohammed Awad and Monir Foqaha combined two techniques of ANN named radial basis function neural networks (RBFNN) and particles swarm optimization(PSO) [14]. This article use PSO algorithm to improve learning algorithm and network of RBFNN algorithm. K-NN and SVD was also used to improve width and weight of RBFNN algorithm that results in better accuracy. [14]

K-NN is considered as one of better approach for email classification. Er. Geetanjli Chawla and Ritu Saini refine K-NN algorithm which turns to be a more time efficient algorithm [15]. The authors improved KNN algorithm by removing recalculations of internal centers and values which improves accuracy and precision. They also compared the Naïve Bayesian and the refined KNN algorithm and results concludes that Refined KNN algorithm gives better results for Email Spam Detection. [15]

Parhat Parveen and Prof. Gambhir Halse experimented different classification techniques on email dataset to find out which classifier is best in spam email filtration [16]. This research concludes that Naïve Bayes classifier gives better accuracy than Decision tree J48 and SVM algorithm so Naïve Bayes is considered as best algorithm for spam email filtration. [16].

Spam emails can cause economical losses for users. Navid Khalilzadeh Sourati and his co-author proposed efficient algorithm for spam filtration with help of ML techniques [17]. With using a multilayer perceptron model, Authors use three ML algorithms to separate spam email from useful emails and to get high efficiency and low error rates. The authors conclude that the proposed system results in high efficiency as compared to NB and DT J48 classifier algorithm as proposed system gives low rate of false positive [17].

As we know People feels comfort to share their feelings on internet. Sentiment analysis is used to analyze user's mood and views etc. vadlamani ravi and co-author presented a study on SA including literature from 2002-2015 related to machine learning and NLP techniques used for SA [18]. This article presents a list of available datasets for Sentiment analysis which is main contribution of it. [18]

Spammer always finds new ways to reach people. many spam filtering method has been introduced but no method will able to remove 100% spam emails. Bayesian is considered as easiest and important approach for filtration of spam emails. Eberhardt, Jeremy J. analyzed two optimizations of Naïve Bayes approach for spam classification i.e. "Multinomial Naive Bayes and Multivariate Naive Bayes" [19]. This article concludes that minor modification in Naïve Bayes algorithm can make significant difference in accuracy as showed in article. [19]

Email classification can be done for several purpose such as subject classification and spam email filtration etc. Izzat Alsmadi and Ikdam Alhami collected dataset of emails and use many approaches to classify the emails basis on their content [20]. They use approaches i.e. SVM, NGram, IDF for email classification and concluded that NGram based clustering give better accuracy but also mentioned that accuracy depends no. of folders in classification technique. [20]

People are typically paying attention to find user's given positive and negative opinions shared for any specific product or service. Feature extraction in SA become active field because it uses to extract features of product reviews. Different techniques have been used for feature extraction in SA. Muhammad Zubair Asghar and his co-author reviews the different existing feature extraction methods for sentiment analysis [21]. They discussed literature related to many machine learning methods, feature selection techniques and clustering base approaches etc. They categorized features into six types such as semantic, syntactic, implicit, frequent, explicit and lexico-structural. They also discussed pre-processing methods i.e. PoS tagging, stopwords removal, stemming & lemmatizing etc. [21] This article concludes that conclude that "feature space reduction, redundancy removal and evaluating performance of hybrid methods of feature selection" can be used as research gap for future research [21].

In this article [22], main goal of authors is to develop a filter which filters spam emails on basis of user preferences with help of known sample.to gain the best efficiency they used TFDCR that can dynamically update feature space. It can also quality of do incremental update in classifier. This study combined SVM with incremental learning which increase classifier's accuracy and performance. [22]

This article [23] presents a hybrid approach for email classification in case of if we have to classify spam and divide into more than two categories. This study combines neural network with SVM. In case of two spam categories neural network will be used for classification but in case of more than two categories first neural network is used then apply SVM according to proposed methodology. This study has main objective to classify emails in to spam and then further divides in to categories. They evaluate their proposed model for four categories by finding its precision and recall and it gives maximum 85% precision and 84% recall. [23]

Many learning algorithm has been observed for purpose of spam email filtration. This article [24] proposed a new algorithm form spam filtration which is named as “antlion optimization (ALO) and boosting termed as ALO-Boosting”. This model has been compared with previous approaches used for spam filtration but proposed model results in better precision. [24]

This article [25] covers commonly used machine learning techniques form spam filtration, important concepts of it. They compared pros and cons of these techniques and discussed basic process of filtration and after the comparison they had made they suggest to use deep learning techniques for future, [25]

3. PROPOSED METHODOLOGY

3.1. Research Type

Our focus is to explore categorize email on basis of feature based sentiment analysis using supervised learning in this research. We will design a model and implementation of that model in a tool in order to perform email categorization.

3.2. Proposed Methodology

The Proposed Methodology consists of steps explained below and shown in figure 1.

3.2.1. Collection of Dataset Abbreviations and Acronyms

In this research, data collected from personal Gmail account having one thousand emails. we have considered body part of email for this research because this research belongs to text classification. After cleaning (spam or junk removing) three hundred email selected for order/command set and three hundred for request set. Three hundred for general set is collected from Enron email dataset. [26]

Each set contains more than 10,000 words. we have done tagging of dataset with labels as request set labeled with 0, order set with 1 and general with 2. [27]

3.2.2. Split Sentence

In Dataset, each email sentence is labeled with tag but we need only sentence part for preprocessing stage so we split our dataset in two categories which are label and sentence.

3.2.3. Preprocessing

Preprocessing is considered as important stage in classification and categorization task as it is important to preprocess the data for extracting accurate meaning from text. Usually text contain those words which make it ambiguous so there is need to clean the text from these types of words to make text meaningful. preprocessing stage is divided into following steps.

- Stop words Removal
- Special Symbol Removal
- Punctuation Removal
- Tokenization

3.2.4. Feature Vector

Vector is series of numbers. Feature vector contains more than one element, put these vector together for creating feature space.in simple words, it contains important characteristics of data. Deep learning work in indexes so we converted our words into indexes like matrices. then vocabulary size is defined according to these indexes. these word2index is stored separately in variable and data is again converted to words for further processing.

3.2.5. Word Embedding

Word Embedding uses word2vec technique which converts words into vector. Google developed group of deep learning models named word2vector which covers context of words. we have imported word2vector file of google for assigning weight to our dataset.

3.2.6. Sentence Creation

In Preprocessing, tokenizer convert sentences into words. to maintain sequence of these words we use pad_sequence function. then we use append function for making sentences from tokenize words and also append their tags which we removed in split stage. these tags were stored in array so we appended them according to their indexes.

3.2.7. Division of Dataset

Deep learning need two portion of data in which one is used for purpose of machine training and other is used for machine testing purpose. we have divided dataset into two part. 90% of dataset is used for machine training and 10% is used for machine testing purpose.

3.2.8. Deep learning Algorithm

Deep learning algorithms are considered as powerful algorithm because it has ability to process huge number of features. We have used convolutional neural networks and recurrent neural network algorithm for text classification.
it performs following steps:

1. Take input from dataset which is reserved for training.
2. Define batch size (collection of data for one cycle) and number or EPOCHS (training cycle).
3. Finally, it trained machine for further computation.

3.2.9. Classifier

After training of machine has been completed. data is tested on the machine then machine acts like classifier and do the classification.

3.2.10. Results

Proposed model gives better results than previous used Fuzzy ANN. It gives accuracy of 0.862 with CNN and 0.949 with RNN.

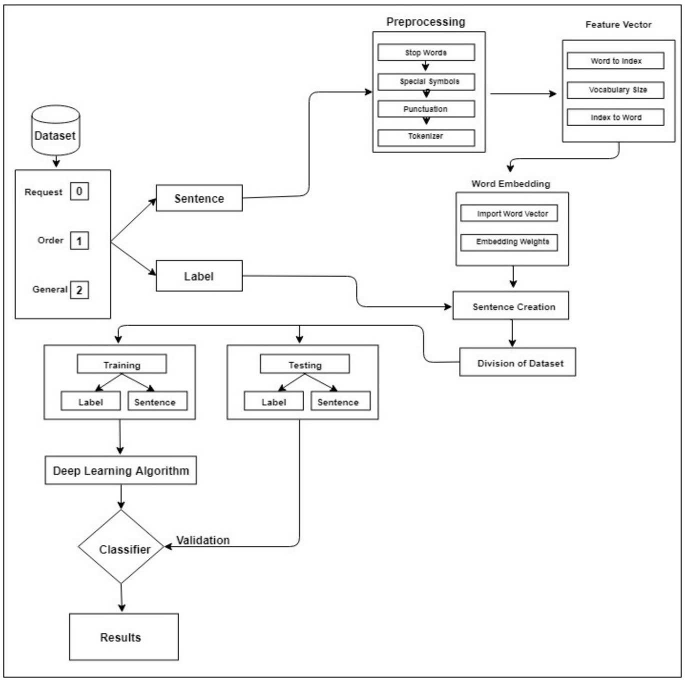


Figure 1. Proposed Methodology

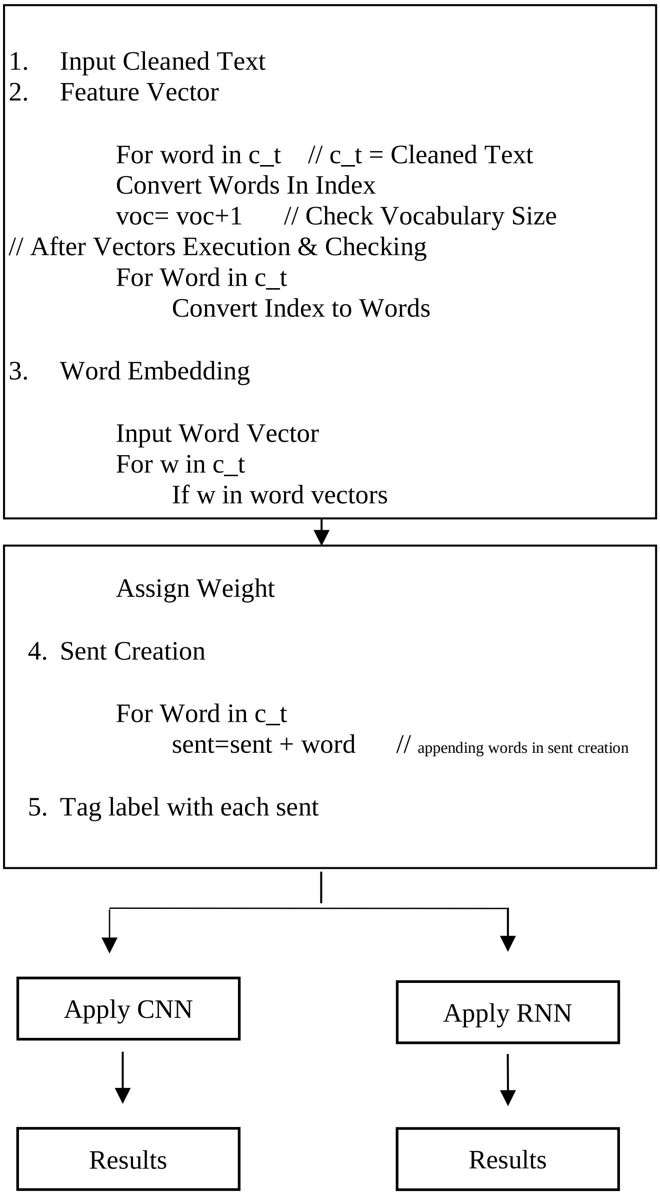
3.3. Model Algorithm Overview

Flow of algorithm is shown in figure is based on following steps.

1. Input layer using word2vector technique for converting words into vectors
2. Split dataset
3. Preprocessing
4. Apply word embedding technique on processed data
5. Appending the processed data
6. Apply CNN and RNN on appended data
7. Final output layer

3.4. Detailed Model Algorithm

```
1.  Input Text
2.  Import vectors              //Meaningful Words
3.  Split
      i.  Sentence
     ii.  Label
4.  Text Preprocessing
    Tokenizer     // Cleaning text

    For i in tokens // i is pointing a word

    If i in stopwords // Removing Stopwords

            remove stopword

    else

             if i is not Meaningful Word

                     Apply Lemmatizer

    else

            Apply Punctuation

5.  Return Cleaned Text
```



```
1.    Input Cleaned Text
2.    Feature Vector

            For word in c_t    // c_t = Cleaned Text
            Convert Words In Index
            voc= voc+1      // Check Vocabulary Size
// After Vectors Execution & Checking
            For Word in c_t
                  Convert Index to Words

3.    Word Embedding

            Input Word Vector
            For w in c_t
                  If w in word vectors

            Assign Weight

  4.  Sent Creation

            For Word in c_t
                  sent=sent + word      // appending words in sent creation

  5.  Tag label with each sent
```

4. Results

We have defined a methodology in order to classify email. Classification of email is done on basis of two features order/command and request. For classification, two deep learning algorithm was used for the testing and training of machine over dataset which are CNN and RNN. We have compared both algorithms on basis of accuracy, precision, recall and f-measure. Some terms we have used in order to find accuracy of models which are “true positive, true negative, false +ve and false -ve”. The result is called true positive when model predicts positive class accurately. When model predicts negative class accurately is called true negative. False positive is incorrectly predicting of positive class where false negative is incorrectly predicting of negative class.

In this research, we have used following formulas for calculating precision, recall and f-measure.

Recall is usually used to find al real positive cases so it Is stated as true positive rate. [28] [29]

$$\text{Recall=true positive rate=TP/RP=TP/(TP+FN)} \qquad (1)$$

Precision is usually known as true positive accuracy because it is used to find how accurately model has find true positive classes. [28] [30]

$$\text{Precision=True Positive Accuracy=TP/PP=TP/(TP+FP)} \qquad (2)$$

F-measure is usually known as mean of precision and recall. [28] [31]

$$\text{F-MEASURE=2-TPR/[TPR+C.FPR+1]} \qquad (3)$$

Formula for finding accuracy is given below. [28] [32]

$$\text{ACCURACY=[TPR+C.(1-FPR)]/[1+C]} \qquad (4)$$

4.1. Results with CNN

we have used CNN for classification. Dataset is trained and tested using CNN, WE get following result using CNN.

Table 1. CNN Results of Email Categorization

MODEL	Accuracy	F-measure	Recall	Precision
CNN	0.862	0.913	0.933	0.921

4.2. Results with RNN

After testing dataset using CNN, we have used RNN to see if we get better result than CNN. Following are results we get by using RNN.

Table 2. RNN Results of Email Categorization

MODEL	Accuracy	F-measure	Recall	Precision
RNN	0.949	0.936	0.908	0.922

4.3. Comparison between RNN and CNN

Following are result we get after comparison of CNN and RNN. we RNN outperform CNN in term of accuracy f-measure and precision but CNN give good result in term of recall as shown in below table and graph.

Table 3. CNN Vs RNN

MODEL	**Accuracy**	**F-measure**	**Recall**
CNN	0.862	0.913	0.933
RNN	0.949	0.936	0.908

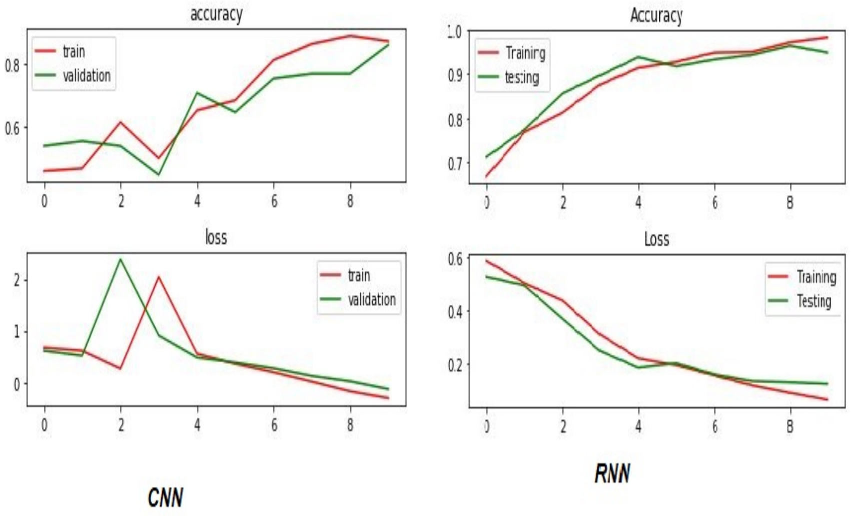


Figure 2. RNN Vs CNN

5. Evaluation

we have used different ratios of dataset used for training and testing as elaborated in table 4 and table 5 by which we have come to conclude that dataset ratio used for training directly proportional to accuracy of algorithm.

Table 4. CNN Evaluation

Train Test Ratio	Accuracy	F-measure	Recall	Precision
90%-10%	0.862	0.913	0.933	0.921
80%-20%	0.815	0.885	0.921	0.902
70%-30%	0.805	0.928	0.822	0.871
60%-40%	0.777	0.857	0.832	0.843
50%-50%	0.802	0.871	0.903	0.884
40%-60%	0.735	0.885	0.760	0.812
30%-70%	0.487	0.551	1.000	0.706
20%-80%	0.682	0.736	0.896	0.803

Table 5. RNN Evaluation

Train Test Ratio	Accuracy	F-measure	Recall	Precision
90%-10%	0.949	0.936	0.908	0.922
80%-20%	0.918	0.895	0.854	0.874
70%-30%	0.903	0.877	0.826	0.850
60%-40%	0.896	0.866	0.815	0.840
50%-50%	0.885	0.844	0.806	0.824
40%-60%	0.862	0.808	0.771	0.789
30%-70%	0.836	0.764	0.736	0.749
20%-80%	0.774	0.667	0.642	0.654

We have done evaluation of both CNN and RNN algorithm From the above comparison, it is concluded that more dataset used for training results in better result and also RNN gives better results than CNN in every case.

Moreover, we have compare our methodology with previously used techniques for email classification in research article [4] as shown in table 6 and figure 3, 4, 5 and 6.

Table 6. Comparison of Deep Learning Techniques

Algorithm	Recall	Precısıon
Fuzzy ANN	86.2	87.3
CNN (Proposed)	93.3	92.1
RNN (Proposed)	90.8	92.2

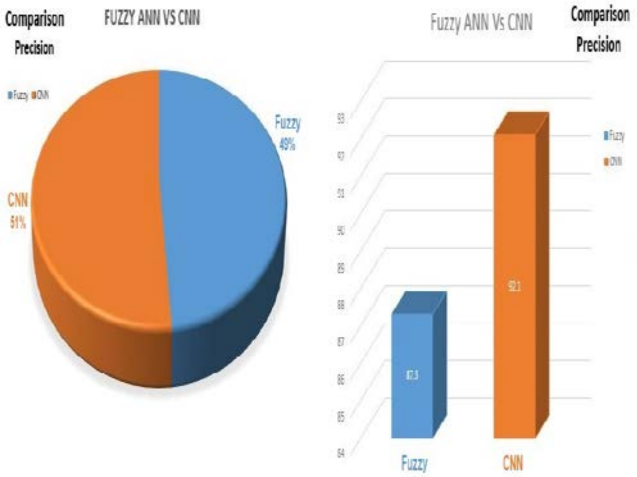


Figure 3. CNN Vs Fuzzy ANN in terms of Precision

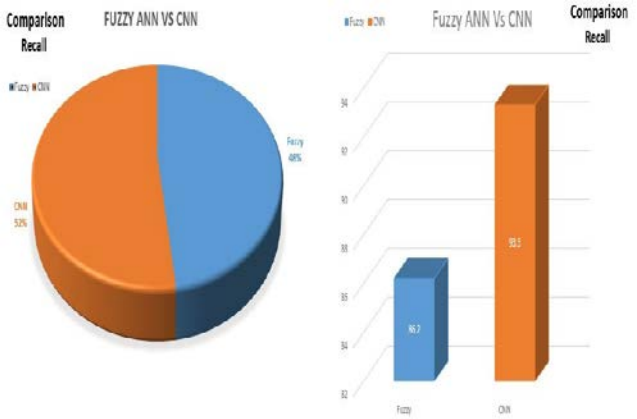


Figure 4. CNN Vs Fuzzy ANN in terms of Recall

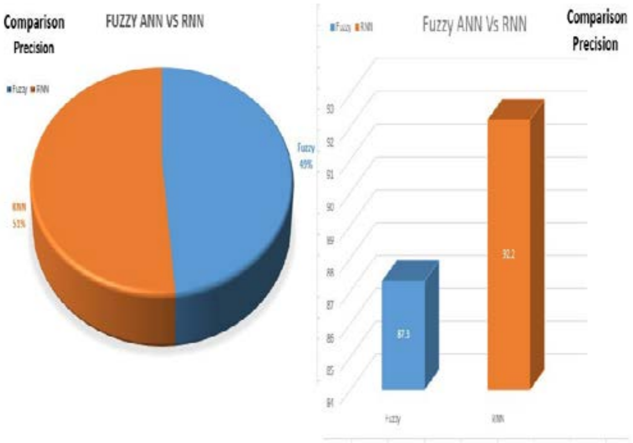


Figure 5. RNN Vs Fuzzy ANN in terms of Precision

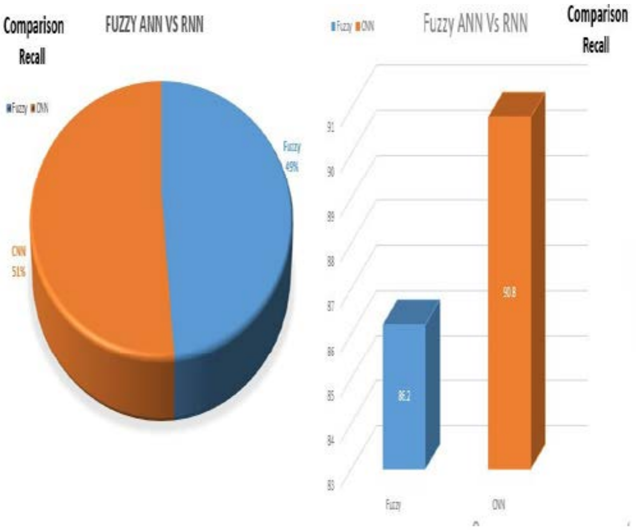


Figure 6. RNN Vs Fuzzy ANN in terms of Recall

5. CONCLUSION

In this research, our focus is to categorize email body text into three categories i.e. request, order/command and general. general category refers to anything other than request and order/command. in this research, we have collected dataset from personal Gmail account except dataset for general category which is collected from Enron. dataset consists of nine hundred e-mail in which three hundred e-mails for each category. From this dataset, 90% dataset is used to train the machine and random 10% is used to test the machine. we have proposed a methodology for email text classification. in this model we use deep learning algorithm as classifier. RNN and CNN are used as deep learning algorithm. then we made a comparison of result of both algorithms. after the comparison of both algorithm, in this research, it is observed that RNN perform better than CNN in term of accuracy as it gives accuracy of 0.949 over 0.862 of CNN. After the comparison of Proposed models with Previously used Fuzzy ANN it is concluded that Our proposed methods CNN and RNN gives better results than Fuzzy ANN in term of Recall and Precision as CNN gives 93.3% Recall and 92.1 precision where RNN gives 90.8% Recall and 92.2 precision which is better than 86.2% Recall and 87.3% precision of Previous Method Fuzzy ANN. This research has also included different experimental result in which CNN and RNN applied on different ratios of training and testing dataset. These experiment show that increasing in the ratio of training dataset results in increasing accuracy of algorithm as CNN and RNN gives highest accuracy of 86% and 94% in case of 90%-10% training-testing ratio and lowest accuracy of 68% and 77% in case of 20%-80% training-testing ratio.

As this research works on classification of email body text in to three features i.e. order, request and general. In future, our focus is to test this model on bigger dataset to find out the accuracy of our model. We also plan to improve accuracy of our model by using other deep learning technique. We believe that using multi-model approach and hybrid approach can improve accuracy. we have also plan to update feature space in future.

REFERENCES

[1] I. THE RADICATI GROUP, "Email Statistics Report, 2016-2020," A TECHNOLOGY MARKET RESEARCH FIRM, USA, 2016.

[2] S. W. a. C. L. Sidner, "Email Overload: Exploring Personal Information formation," CHI, pp. 276-283, 1996.

[3] L. S. R. G. R. M. A. M. A. A.-G. GHULAM MUJTABA, "Email Classification Research Trends: Review and Open Issues," Digital Object Identifier, 2017.

[4] A. N. K. M. K. G. P. G. Akash Kumar Singh, "Email Classification using NLP & Machine Learning Techniques," IJSRD - International Journal for Scientific Research & Development, vol. 6, no. 3, 2018.

[5] A. A. T. I. J. S. B. &. J. A. Hanif Bhuiyan, "A Survey of Existing E-Mail Spam Filtering Methods Considering Machine Learning Techniques," Global Journal of Computer Science and Technology, vol. 18, no. 2, pp. 0975-4172, 2018.

[6] D. R. S. R Manikandan, "Machine learning algorithms for text-documents classification: A review," International Journal of Academic Research and Development, vol. 3, no. 2, pp. 384-389, 2018.

[7] S. L. &. I. Lee, "Email Sentiment Analysis Through k-Means Labeling and Support Vector Machine Classification," Cybernetics and Systems, vol. 49, no. 3, pp. 181-199, 2018.

[8] P. K. B. ESHA BANSAL, "A SURVEY OF VARIOUS MACHINE LEARNING ALGORITHMS ON EMAIL SPAMMING," International Journal

of Advances in Electronics and Computer Science, vol. 4, no. 3, pp. 2393-2835, 2017.

[9] P. V. a. S. R. D. Karthika Renuka, "An Ensembled Classifier for Email Spam Classification in Hadoop Environment," Applied Mathematics & Information Sciences, vol. 11, no. 4, pp. 1123-1128, 2017.

[10] P. T. V. Shradhanjali, "E-Mail Spam Detection and Classification Using SVM and Feature Extraction," International Journal of Advance Research, Ideas and Innovations in Technology, vol. 3, no. 3, pp. 1491-1495, 2017.

[11] O. L. W. O. A. M. A. Abubakr H. Ombabi, "Deep Learning Framework based on Word2Vec and CNN for Users Interests Classification," in Sudan Conference on Computer Science and Information Technology (SCCSIT), 2017.

[12] D. H. A. Priti Kulkarni, "Comparative analysis of classifiers for header based emails classification using supervised learning," International Research Journal of Engineering and Technology (IRJET), vol. 3, no. 3, pp. 1939-1944, 2016.

[13] A. G. d. P. V. S. L. F. G. C. Rogerio Bonatti, "Effect of Part-of-Speech and Lemmatization Filtering in Email Classification for Automatic Reply," in The Workshops of the Thirtieth AAAI Conference on Artificial Intelligence Knowledge Extraction from Text:, 2016.

[14] M. A. a. M. Foqaha, "EMAIL SPAM CLASSIFICATION USING HYBRID APPROACH OF RBF NEURAL NETWORK AND PARTICLE SWARM OPTIMIZATION," International Journal of Network Security & Its Applications (IJNSA), vol. 8, no. 4, pp. 17-28, 2016.

[15] E. G. C. a. R. Saini, "Implementation of Improved KNN algorithm for Email Spam Detection," International Journal of Trend in Research and Development, vol. 3, no. 5, pp. 479-483, 2016.

[16] P. G. H. Parhat Parveen, "Spam Mail Detection using Classification," International Journal of Advanced Research in Computer and Communication Engineering, vol. 6, no. 6, pp. 347-349, 2016.

[17] N. K. S. Ali Shafigh Aski a, "Proposed efficient algorithm to filter spam using machine learning techniques," Pacific Science Review A: Natural Science and Engineering, pp. 145-149, 2016.

[18] V. R. Kumar Ravi, "A survey on opinion mining and sentiment analysis: Tasks, approaches and applications," Knowledge-Based Systems, pp. 14-46, 2015.

[19] J. J. Eberhardt, "Bayesian Spam Detection," Scholarly Horizons: University of Minnesota, Morris Undergraduate Journal, vol. 2, no. 1, 2015.

[20] I. A. Izzat Alsmadi, "Clustering and classification of email contents," Journal of King Saud University –Computer and Information Sciences, pp. 46-57, 2015.

[21] A. K. S. A. F. M. K. Muhammad Zubair Asghar, "A Review of Feature Extraction in Sentiment Analysis," Journal of Basic and Applied Scientific Research, 2014.

[22] K. K. Gopi Sanghani, "Incremental personalized E-mail spam filter using novel TFDCR feature selection with dynamic feature update," Expert Systems With Applications, vol. 115, p. 287–299, 2019.

[23] S. Sanjeev Dhawan, "An enhanced mechanism of spam and category detection

using Neuro-SVM," Procedia Computer Science , vol. 132 , p. 429–436, 2018.

[24] N. I. G. ,. A. A. S. Amany A. Naem, "Antlion optimization and boosting classifier for spam email detection," Future Computing and Informatics Journal , vol. 3, pp. 436-442, 2018.

[25] J. S. B. ,. H. C. ,. M. A. ,. A. O. A. ,. O. E. A. Emmanuel Gbenga Dada, "Machine learning for email spam filtering: review, approaches and open research problems," Heliyon, vol. 5, 2019.

[26] O. Halvani, " "Enron Authorship Verification Corpus", Mendeley Data, v2," (2018. [Online].
Available: https://data.mendeley.com/datasets/n77w7mygwg/2/files/4c220a60-b725-4c58-ac12-9c81b0300bce.

[27] M. T. O. Reshmi Gopalakrishna Pillai, "Detection of Stress and Relaxation Magnitudes for Tweets," The Sixth International Workshop on Natural Language Processing for Social Media, pp. 23-27, 2018.

[28] D. POWERS, "EVALUATION: FROM PRECISION, RECALL AND F-MEASURE TO ROC, INFORMEDNESS, MARKEDNESS & CORRELATION," Journal of Machine Learning Technologies, vol. 2, no. 1, pp. 37-63, 2011.

[29] A. &. D. M. Fraser, Computational Linguistics, 2007.

[30] Reeker, "Performance Metrics for Intelligent Systems," 2007.

[31] Carletta, "Computational Linguistics," 1996, pp. 249-254.

[32] Hutchinson, Research in Nursing & Health, (1993).