HIERARCHICAL TEXT CLASSIFICATION USING DICTIONARY-BASED APPROACH AND CONVOLUTIONAL NEURAL NETWORK

Ms. Sneha N. Jagtap
M. Tech Student
Dept. of Computer Science and Engineering
Rajarambapu Institute of Technology, Sakharale, India.
jagtap.sneha35@gmail.com

Amol. C. Adamuthe
Assistant Professor, Dept. of Information Technology
Rajarambapu Institute of Technology, Sakharale, India.
amol.admuthe@gmail.com

***Abstract*— As increasing amount of electronic information which is usually in textual form. There is an important requirement to classify textual reports generated by organizations into various categories. Machine Learning techniques can be employed to classify text autonomously in an effective manner. We propose a generic web-based application for classifying text using machine learning techniques. For creating a baseline implementation, we projected a dictionary-based approach for text classification. The dictionary is populated based on domain-specific inputs. The dictionary is expanded using automated techniques. This helps us in getting an end-to-end implementation of full pipeline. In subsequent phases, we proposed our approach with Convolutional Neural Network (CNN) for text classification.**

Keywords—Machine learning, Text mining, Text Classification, Dictionary-Based Approach, Convolutional Neural Network,

I. Introduction

Text mining research has become more and more important recently because of the rapidly growing up number of electronic documents from various sources. Semi-structured and unstructured resource information includes the government electronic information databases, World Wide Web, biological databases, news articles, online forums, digital libraries, chat rooms, email and blog libraries. So, particular knowledge discovery and classification from large resources is a most important research area. data mining, natural language processing as well as machine learning techniques are work together to discover patterns and automatically classify in electronic documents. The important aim is to allow clients to handle such operations as search, classification and abstract and extract information from text sources and [1].

Categorization of text is a categorization of text about a group of one or more earlier existing categories. Classification of text is a very useful operation, often used to assign subject categories to documents, to route and filter text, or as part of a natural language processing system [2]. Text classification is the division of a group of input text into two or more classes, each of which can be said to belong to one or more classes.

In the classification system, a set of words or terms are collected and organized. Each of these terms will be associated with a specific concept. Classification systems are usually hierarchical, which means that more details get additional details in the next level, and concepts are related to and planned around each other's characteristics [3].

Machine learning itself is a vast field of computer science. Text categorization is part of a wide range of machine learning methods. This classification problem can be solved by various algorithms. In the work of this proposal, dictionary-based methods and convolutional neural networks were used to classify text documents. Convolutional neural networks, in short, ConvNets are advantageous in computer views. The important concept of ConvNets is to deal with classification and feature extraction as a well-trained task. This idea has been improving over the years, especially by using hierarchical convolutions and pools to sequentially extract the hierarchical representation of the input [4].

In this paper, section 2 gives literature survey on text classification and proposed methodology, section 3 explains problem formulation, section 4 explores a proposed methodology and explanation on it and the last section shows implementation results on proposed methodology.

II. Literature review

A study of existing theories and practices (literature) in the chosen area or domain helps in identification of gaps or deficiencies in knowledge and in scoping the study by identification of limitations and assumptions. All this helps in framing the problem statement.

A. *Literature Review: Text classification*

In paper [5] Aurangzeb Khan *et al.* the intention is to focus on useful methodologies and techniques which are used in documents classification. Meanwhile some of the challenges that remain to be resolved and also highlighted on machine learning approaches and text representation. This paper gives a view of the methods as well as theory of document classification Also text mining technique highlighting on the current literature.

In paper [6] M. Ikonomakis *et al.* illustrates classification of text process using machine learning approaches. For any supervised learning technique tasks, firstly corpus is required. If any document from the corpus is classified into greater than one category then authors said that this type of problem is solved by ranking classification. They proposed processes of text classification such as stemming, tokenizer, removing stop-words, feature selection, and representation of the vector.

They gave various different metrics for feature selection which are TFIDF, Information Gain, Chi-square etc.

In paper [7] Zakaria Elberrichi *et al.* investigate procedure that use wordnet dictionary concept to classify documents. For that purpose, they use wordnet dictionary which contains synsets. There are two corpora used such as Reuters-21578 and 20 newsgroups for categorization of text. This technique selects a common concept from the dictionary and after that merge in such a way that form a new vector representation. This proposed method is useful for increasing F1value. For dataset, Reuters increased from 0.649 to 0.714 and for 20 newsgroup dataset increased from 0.667 to 0.719.

In paper [8] Vandana korde *et al.* explained survey on classifiers as well as text classification. This paper analyzed methods for feature selection and algorithms were presented. The author gives compressed information to various representations for text. This author did a survey on Roccio's algorithm, KNN, Naïve Baye's, decision rule, decision tree, SVM, Neural Network, LLSF etc. classifiers and focused on literature review.

B. *Literature Review: Text Classification other than Dictionary-Based Approach*

In paper [9] Zulfany Erlisa Rasjid *et al.* focus on data classification using two out of the six approaches of data classification, which is Naïve Bayes and k-Nearest Neighbors. The text documents used is in XML format. The Corpus used in this research is downloaded from TREC Legal Track with a total of more than three thousand text documents and over twenty types of classifications. Out of the twenty types of classifications, six are chosen with the most number of text documents. The data is processed using RapidMiner software and the result shows that the optimum value for k in k-NN occurs at k=13. Using this value for k, the accuracy in average reached 55.17 percent, which is better than using Naïve Bayes which is 39.01 percent.

In paper [10] Mark Hughes et al. illustrate a method to certainly classify at a sentence level dataset which is present clinical text. The deep convolutional neural network used for the representation of complex features. They are providing dataset immense information of health classification. Through detailed assessments, they demonstrated that their method outweighs about 15% of the many methods used in tasks of natural language processing.

In paper [11] Xiang Zhu *et al.* maintain a new technique depends on summarization of text techniques, convolutional networks, and word vector model. Text classification is classified by using a convolutional neural network. This technique is first selects summary of given social networking articles and after that by using word vector model that summary is converted into vectors.

In paper [12] Shiyao Wang *et al.* established a new methodology or classification of text termed as DIST.They used AG news corpus for text classification and results show 7.99% test errors which have the best result ever.

C. *Literature Review: Convolutional Neural Network*

In paper [13] Guibin Chen et al. propose an application of recurrent neural networks and convolutional. This proposed method used to catch both the local and the global semantics or textual as well as to perfect high-order label correlations while having a tractable computational complexity.

In paper [14] Manabu Nii et al. illustrate CNN-based classification technique for evaluating nursing-care texts. Every nursing-care text is represented as a vector which concatenated word vectors. Word vectors are obtained by using the word2vec.The proposed CNN-based method gives us better classification results for nursing-care text evaluation than our previous works. For each subset of the nursing-care data, the proposed CNN-based method shows not always better results.

In paper [15] Xiang Bai et al. projected a new convolutional neural network architecture which is termed as MSP-Net. This architecture is used for classification of text as well as non-text images. Input is taken by full image and output is produced in an end-to-end manner of classification of block-level.

D. *Literature Review: Text Preprocessing*

In paper [16] Alper Kursat Uysal *et al.* gives information about the impact of pre-processing on a classification of text. This paper having an intention to widely analyze the pre-processing impact on classification of text. This impact is having various features such as reduction of dimensions, a language of a text, text domain, accuracy of classification. Pre-processing techniques are implemented on two different domains like news as well as email and two different languages like English and also Turkish.

In paper [17] Tajinder Singh *et al.* proposed work is to analyze the impact of pre-processing and normalization used for small messages filled with many features like symbols, information, noise, fencoloric, abbreviations, and unspecified words. The proposed scheme used in this paper first gathers the coexisting words with the slang and then exploits characteristics of these binding words to define the significance and sentiment strength of slang word used in the tweet which not only facilitates the better sentiment classification but also ensures the sturdiness of classier as shown in the results.

In paper [18] Dina A. Said, et al. proposed an Arabic Text Categorization by using morphological tools. The proposed work involves using raw text, stemmed text, as well as a root text. With the help of pre-processing tools root text and stemmed are obtained.

III. PROBLEM FORMULATION

In today's digital world, all information is transformed into a digital form which is presented into textual form. There is an important requirement to classify textual reports generated by organizations into various categories. Machine Learning techniques can be employed to classify text autonomously in an effective manner. It requires less time for searching any document which is required. Text classification system could be helpful in classifying emails which are spam or ham, classifying news which is whether a political or sports-related news, classifying whether a report is sales related, a complaint related or legal related etc.

We propose a generic web-based application for classifying text using machine learning techniques. We propose two methods for classification of a text document. The first method is classifying document on the bases of dictionary-based approach. For that input is required in the form of a textual file. That file is not containing any image, audio and video contents. By using Dictionary-Based Approach get a correct output such as that file is classified into a particular class. The second method is classifying document by using a convolutional neural network. The traditional convolutional neural network takes an image as input but here we are passing text as input and we are getting output correctly.

IV. Methodology and discussion

Text classification is done by using two algorithmic approaches such as Dictionary-Based Approach and Convolutional Neural Network.

A. Dictionary-Based Approach

Categorization of text is a system that set narrative into one or more existing groups depend on their elements which are present in that file. Classification of text has broad applications, such as filtering of emails, classification of category for search engines as well as digital libraries. The first step that is, 'Pre-processing' is the important subpart of text classification [19]. Text preprocessing system consists of activities like tokenization, stop word removal, stemming. Figure 1 shows Text Classification Process.

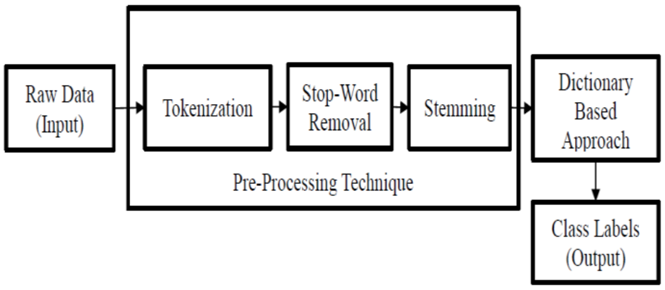


Figure 1: Text classification process using Dictionary-Based Approach

- **Raw Data (Input):**

In this step, raw data contains dataset which is used as input for further process of text classification. Input is present in the form of a text file which does not contain any type of image file, audio, and video file.

- **Text Pre-processing Techniques:**

The primary process is a step in the development of Natural Language Processing, Information Retrieval as well as Text Mining. Pre-processing is needed for selecting interesting, important and informative from unstructured data in text mining domain [19]. Text pre-processing system consists of activities which are vector representation, trimming of stop-word, tokenization as well as stemming.

The purpose behind pre-processing is to enact each document file as a feature vector which is to splits text i.e. sentences into single words. In designed classifiers, text documents are formed as transactions. Selecting the keyword which is the feature selection process, is the vital pre-processing footprint essential for the indexing of documents. This level of stage decides the status of next stage, which is very important for rating the stage. It is very essential to collect useful keywords that have sense and rejects the unwanted texts that didn't donate to the difference amongst the documents [19].

- Tokenization Technique

Tokenization is most important part of text pre-processing techniques. Tokenization is the process of segmenting text into various words, phrases which are presented in text, many of symbols, or other interesting elements (called tokens). The tokenization techniques are defined as the investigation of words which are presented in a sentence. In the proposed system, tokens of sentences are separated by using space. The list of tokens is used for further processes as input. In further processes, various operations are applied on that selected tokens. Tokenization technique is beneficial for both in computer science, where it forms part of the lexical analysis and in semantic (where semantic is in the form of segmentation of text.). The vital benefit of tokenization is recognizing meaningful keywords. Textual information is a chunk of characters which are at the origin. All steps in a retrieval of information need the words of the dataset. Thus, tokenization of documents is the requirement for a parser.

- Removal of Stop-Word Technique

Stop-word removal which is used to erase the words which does not have any meaning and if they are not present in file meaning is not changed. Stop words contains 'and', 'are', 'this' words etc. Those words are not having use in text document classification. Therefore, these words must be erased. The formulation of stop-word vocabulary is crucial and uncertain between text sources. This method erases text information and raises the performance of a system. Each text document handles tokens that are unimportant by these text mining applications [20].

- Stemming Technique

Stemming is one of the pre-processing techniques which is used for finding root word from various words. For example, the words: "tokens", "tokenizer", "tokenization", "tokenizing" all these words are reduced to common root word representation “token”. The determination of root or stem is done by this method. This process is used to eliminate different suffixes, for trimming the number of words, to match the perfect stems, and also used to reduce time and storage space [20].

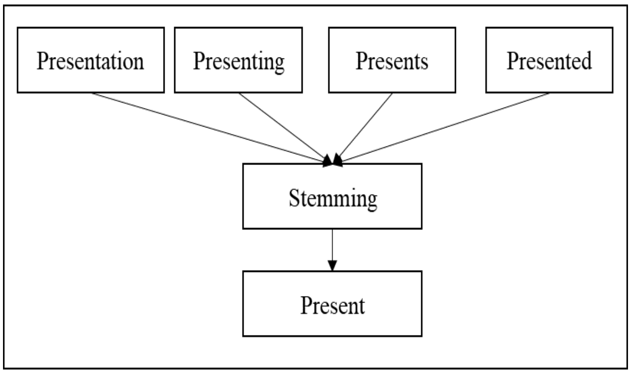


Figure 2: Stemming process

Algorithms of stemming methods are divided into three parts: mixed methods, truncating methods as well as statistical methods. Mixed method is divided into Corpus-based, Inflectional and Derivational a) Xerox b) Krovitz. Statistical stemmer is divided into three subparts such as YASS, HMM, N-Gram and last type of Stemming algorithm truncating are divided into further four parts such as Lovins, Dawson, Paice/ Husk and porter [21]. In the proposed system, porter stemmer algorithm which is sub part of truncating stemming algorithm is used for stemming purpose. It is one of the best stemmer algorithms for finding root word from the files. Working o porter stemmer algorithm on one example is as follows such as OSCILLATORS. Now we will get OSCILLATOR from 1st step of the algorithm, then get OSCILLATE in 2nd step of stemmer algorithm and then get OSCILL in 3rd step of an algorithm and then last get OSCIL in 4th step of stemming algorithm.

- **Dictionary-Based Approach:**

Dictionary-Based Approach is most important part of text classification. Pre-processing techniques are performed on input files and output of that file are given to the Dictionary-Based Approach as input. But Before that input file is converted into a vector form of a file with the help of file which is generated from a global list of dictionaries. After that dot product on it. Then the file is classified into proper predefined classes.

Flowchart:

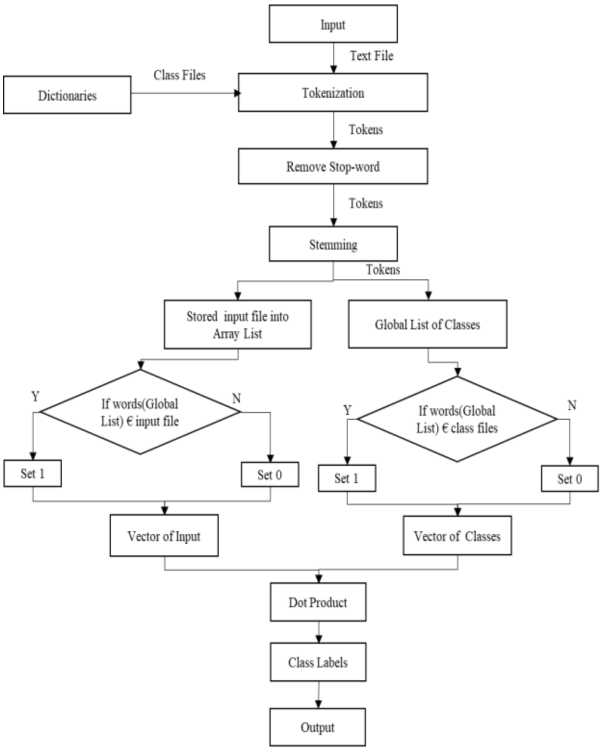


Figure 3: Flow Chart of Dictionary-Based Approach for Text Classification

Pseudo Code:

```
1. Take a text file which does not contain image, video and audio content in that file as input
2. Split text into small tokens such as words, phrases, symbols etc.
3. Delete words which are not having meaning if removed from that file.
4. Find stem words from an input.
5. Store tokens into an array list of input file
6. Perform step 2,3 and 4 on dictionary files
7. Make a global list from given dictionaries.
8. Convert given file into vector form by using global list.
a. If (word (Global List) € input file)? 1: 0
9. Convert predefined classes into a vector form with the help of the global list.
a. If (word (Global List) € class file)? 1: 0
10. Perform Dot Product on those vector files.
11. After that file is classified into a particular class.
```

- **Class Label (Output):**

Class Labels means an output of given dataset. Here, a Particular file is classified as its particular given label.

B. Convolutional Neural Network

Convolutional Neural Network is another approach which is used in proposed methodology for text classification. Neural Networks is same as a human brain. Convolutional Neural Network contains three most important layers such as a hidden layer, output layer, and input layer. Input layer refers to an input of the proposed system. Output layer contains class labels of given dataset. Usually, Input and also output layer is fixed but the hidden layer is not fixed. It depends on datasets. Hidden layer depends on a fully connected layer, convolutional layer, max pooling layer.

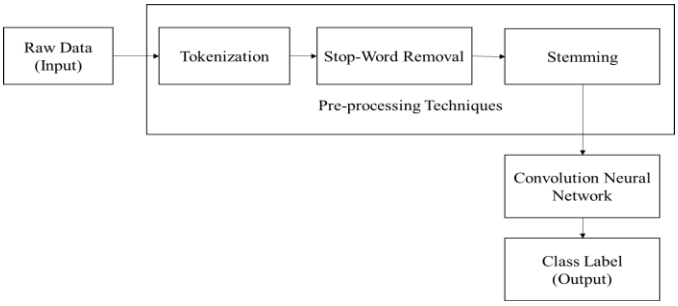


Figure 4: Flow diagram of Convolutional Neural Network for Text classification

The tokenization, stemming and stop-word removal performs an operation by convolution neural network on input files, while running the system. Figure 4 is representing the process of text classification in a convolutional neural network.

Pseudocode:

```
1. Define the data parameters.
2. Model those parameters.
3. Load and save the data.
a. Performed stop word elimination and clean data (remove special characters).
4. Save parameters to file.
5. Performed cross-validation of data (used k-fold cross-validation).
6. Generate mini-batches for training a neural network.
7. Train data for a classifier.
8. Use CNN classifier.
9. Repeat step 5 to 8 until the end of training data.
```

Most rarely convolutional neural network is taking an input as an image. Hence, first, we have to convert text into an image for input purpose. Character quantization is used for converting subsequence of encoded characters as input. Quantization is performed with the help of one hot encoding technique. After encoding of those files, convert string vector into an image which is supply to this method. The text is classified on a character level with the help of convolutional neural network classifier. It contains text file as an input layer. Hidden layer is sub-divided into convolutional, max-pooling as well as a fully connected layer.

Flowchart:

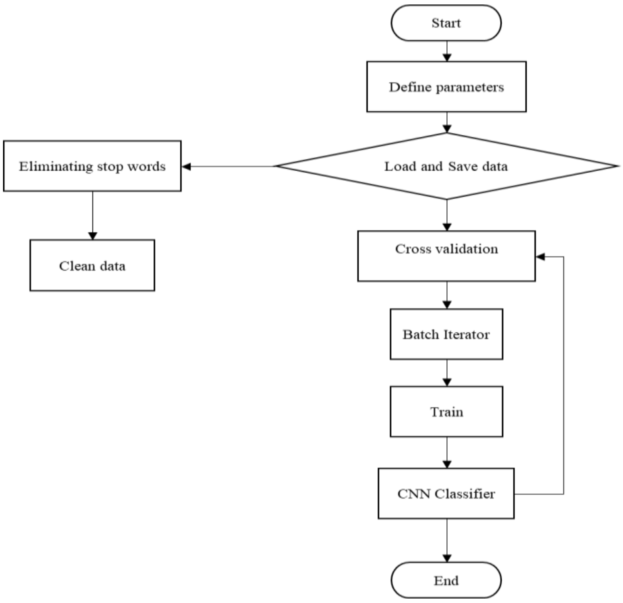


Figure 5: Flowchart of Convolutional Neural Network for text classification.

- Convolutional Layer

A convolutional layer is the main component of a Convolutional neural network which is having local connections as well as weights of characteristics of shared. To grasp feature illustration of given inputs is a main aim of convolutional layer. The objective of a convolutional layer is to extract features of input volume. An output of this layer is feature map or Activation map or convolved feature. A product is applied to image pixels and filter or kernel feature detector, from that we get a convolved feature or activation map or feature map as an output of this layer. In that way, we get different activation map by applying different kernels.

- Pooling Layer

A process of sampling is similar to filtering of fuzzy. Pooling layer is kind of feature extraction which can decrease convolved feature's dimensions and raise the strength of extraction of features. Pooling layer is implanted in between two convolutional layers. The size of activation maps which are presented in pooling layer is decided to depend on moving steps of kernels. Normally pooling layer operations are max-pooling and average pooling. This layer can extract a high characteristic of input.

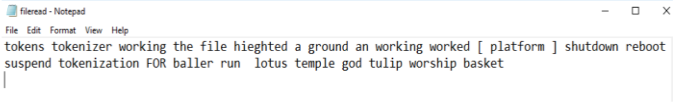
- Fully-Connected Layer

The convolutional neural network contains totally connected layers, that depend on a dataset. Fully connected layer takes all previous layer's neurons and compares them to each one neuron which is presented in a layer. The fully connected layer does not contain any spatial information which is in the form of preserved. An output layer follows a fully connected layer.

V. RESULTS OF TEXT CLASSIFICATION

- Dictionary-Based Approach

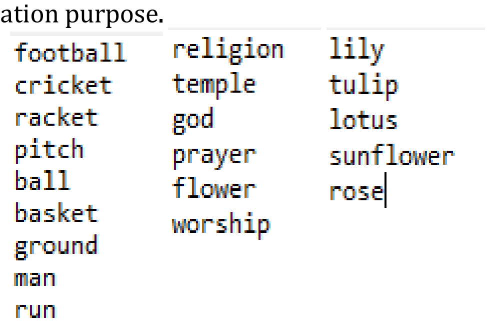
For Implementation of Dictionary-Based Approach, we are taken a file as input which is shown figure 6. This does not contain any images.



```
tokens tokenizer working the file hieghted a ground an working worked [ platform ] shutdown reboot
suspend tokenization FOR baller run  lotus temple god tulip worship basket
```

Figure 6: Input File for Text Classification

Dictionary is required for classification purpose. It contains three classes or dictionaries which belong to sports class, temple class, and flower class. Figure 7 shows three classes which we are used as dictionaries for text classification purpose.



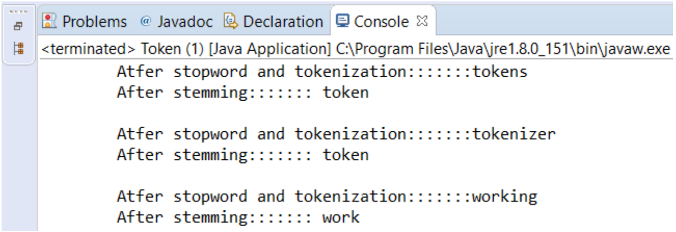
```
football
cricket
racket
pitch
ball
basket
ground
man
run
```

```
religion
temple
god
prayer
flower
worship
```

```
lily
tulip
lotus
sunflower
rose
```

Figure 7: Class files for Text Classification

We apply the pre-processing technique to that input file. An output of this technique is shown in figure 8.



```
<terminated> Token (1) [Java Application] C:\Program Files\Java\jre1.8.0_151\bin\javaw.exe
        Atfer stopword and tokenization:::::::tokens
        After stemming::::::: token

        Atfer stopword and tokenization:::::::tokenizer
        After stemming::::::: token

        Atfer stopword and tokenization:::::::working
        After stemming::::::: work
```

Figure 8: Output after Pre-Processing Techniques on input file

Same operations are performed on all other class files. After that applied dot product for classification. Input text file is classified as sports class.

- Convolutional Neural Network

For implementation of Convolutional Neural Network, we take small dataset which contains two labels and in CSV format. Time is required for a run neural network is approximate 2.14 hrs.

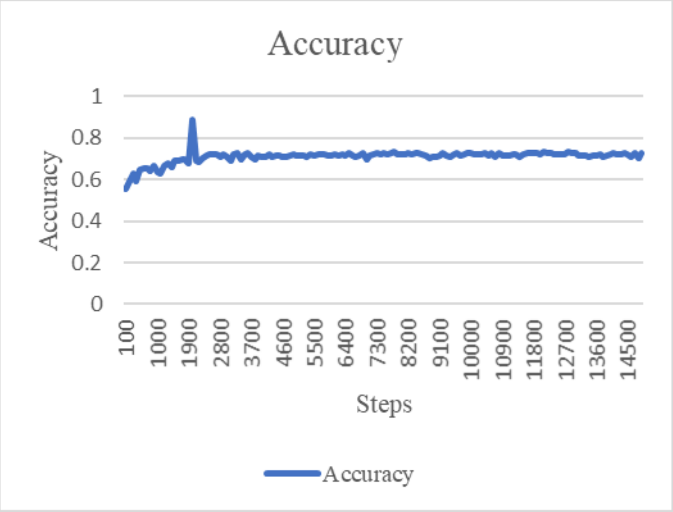


Figure 9: Output of Convolutional Neural Network

This dataset contains 10662 number of sentences. Batch size is 32, maximum document length is 56 and vocabulary size is 18758. It contains 14900 steps iterations. Each contains 32 batch size. There is 50 number of epochs. After applying neural network on given dataset, we are getting accuracy which is shown in figure 9. This figure shows an accuracy on validation which starts slowly increasing at certain point accuracy become constant. Accuracy always lies in between 0 to 1.

VI. CONCLUSIONS

Text classification is an important system for many organizations. This proposed system is useful for maintaining text file in a systematic manner. In this phase, text classification is done by using Dictionary-Based Approach. By using Dictionary-Based Approach, we get the accurate classification of a text file. This will help us in getting an end-to-end implementation of full pipeline. We implement the second approach is Convolution Neural Network for text classification. This approach also gives correct accuracy.

REFERENCES

[1] Hughes, Mark, et al. “Medical Text Classification using Convolutional Neural Networks.” *arXiv preprint arXiv:1704.06841*, 2017.

[2] Vasa, Krina. “Text Classification through Statistical and Machine Learning Methods: A Survey.” 2016.

[3] Uysal, Alper Kursat, and Serkan Gunal. “The impact of preprocessing on text classification." *Information Processing & Management 50.1*, pp. 104-112, 2014.

[4] Khan, Aurangzeb, et al. “A review of machine learning algorithms for text-documents classification.” *Journal of advances in information technology 1.1*, pp. 4-20, 2010.

[5] Ikonomakis, M., Sotiris Kotsiantis, and V. Tampakas. “Text classification using machine learning techniques.” *WSEAS transactions on computers 4.8*, pp. 966-974, 2005.

[6] Elberrichi, Zakaria, Abdelattif Rahmoun, and Mohamed Amine Bentaalah. “Using WordNet for Text Categorization.” *International Arab Journal of Information Technology (IAJIT)5.1*, 2008.

[7] Korde, Vandana, and C. Namrata Mahender. "Text classification and classifiers: A survey." *International Journal of Artificial Intelligence & Applications 3.2* (2012): 85.

[8] Rasjid, Zulfany Erlisa, and Reina Setiawan. "Performance Comparison and Optimization of Text Document Classification using k-NN and Naïve Bayes Classification Techniques." *Procedia Computer Science* 116, pp. 107-112, 2017.

[9] Hughes, Mark, et al. "Medical Text Classification using Convolutional Neural Networks." *arXiv preprint arXiv:1704.06841*, 2017.

[10] Zhu, Xiang, et al. "Chinese Article Classification Oriented to Social Network Based on Convolutional Neural Networks." *Data Science in Cyberspace (DSC), IEEE International Conference on. IEEE*, 2016.

[11] Wang, Shiyao, and Zhidong Deng. "Tightly-coupled convolutional neural network with spatial-temporal memory for text classification." *Neural Networks (IJCNN), International Joint Conference on. IEEE*, 2017.

[12] Chen, Guibin, et al. "Ensemble application of convolutional and recurrent neural networks for multi-label text categorization." *Neural Networks (IJCNN), International Joint Conference on. IEEE*, 2017.

[13] Nii, Manabu, et al. "Nursing-care text classification using word vector representation and convolutional neural networks." *Fuzzy Systems Association and 9th International Conference on Soft Computing and Intelligent Systems (IFSA-SCIS), Joint 17th World Congress of International. IEEE*, 2017.

[14] Uysal, Alper Kursat, and Serkan Gunal. "The impact of preprocessing on text classification." *Information Processing & Management* 50.1, pp. 104-112, 2014.

[15] Bai, Xiang, et al. "Text/non-text image classification in the wild with convolutional neural networks." *Pattern Recognition* 66 (2017): 437-446.

[16] Singh, Tajinder, and Madhu Kumari. "Role of text pre-processing in Twitter sentiment analysis." *Procedia Computer Science* 89 (2016): 549-554.

[17] Vijayarani, S., Ms. J. Ilamathi, and Ms. Nithya. "Preprocessing techniques for text mining-an overview." *International Journal of Computer Science & Communication Networks* 5.1 (2015): 7-16.

[18] Said, Dina, et al. "A study of text preprocessing tools for Arabic text categorization." *The second international conference on the Arabic language.* 2009.

[19] Blumenstein, Michael, Chun Ki Cheng, and Xin Yu Liu. "New preprocessing techniques for handwritten word recognition." *Proceedings of the Second IASTED International Conference on Visualization, Imaging and Image Processing (VIIP 2002), ACTA Press, Calgary*. 2002.

[20] Srividhya, V., and R. Anitha. "Evaluating preprocessing techniques in text categorization." *International journal of computer science and application* 47.11 (2010): 49-51.

[21] Kannan, S., and Vairaprakash Gurusamy. "Preprocessing Techniques for Text Mining." (2014).