Document Classification

Bhargavi G
Computer Science and Engineering
PES University
Bangalore, India
bhargu2000@gmail.com

Abhilash V
Computer Science and Engineering
PES University
Bangalore, India
vabhilash2000@gmail.com

Hari Chandan S
Computer Science and Engineering
PES University
Bangalore, India
harichandan0297@gmail.com

***Abstract*—Handling large volumes or incoming data from various sources manually is highly inefficient and is more prone to human-level errors that can produce incorrect or unexpected results. Automated Document Classification comes under the Natural Language Processing Domain where an algorithm assigns fixed categories to a cluster of textual Documents such as articles,emails etc. Machine Learning approaches are faster, more scalable and less biased when compared to manual classification. In our implementation, we will be using the supervised method where we tag our training documents with specific categories. Our algorithm will learn the associations between the text of each document and the tags it has been assigned to during the training phase. For testing, we will use documents not seen by the algorithm but they must belong to the fixed number of categories. The task of our algorithm is to correctly classify these documents into the correct categories. Examples of such algorithms include Naive Bayes, Support Vector Machines and Neural Networks.**

***Index Terms*—Document, Machine Learning, Recurrent Neural Networks, text classification**

I. Introduction

Our project revolves around solving the problem statement of Document Classification. It is the act of labelling documents into well defined categories according to their content using Machine Learning approaches with the main aim of automating the whole process which is mainly advantageous in handling large volumes of data where the process can be slow and monotonous.The scope of classification of text is at document level, paragraph level, sentence level and sub sentence level. In some applications, we assign more than one class to a single document and in other applications, each paragraph or sentence is assigned a category. Sentence level classification is known to be more granular where the sentence level context is determined by the sentences surrounding it. Here, The memory component becomes an important part of the algorithm and complex tools such as LSTMs are used to identify and handling the underlying context of the sentence. In Document level we accept the syntax and semantics of the sentences to perform the classification.

II. Applications/Use Cases

- Today, almost all data received and handled by various sectors is unstructured which means finding useful insights is very tedious and time consuming.
- Document Classification is used extensively in Software Companies where a specific tag is assigned to a document and sent to a relevant team. It is used in error detection where the algorithm assigns the tag '¡bug¿' to a document after analysing it's content and can be routed to the corresponding team.
- A faster emergency response system can be developed by differentiating between 'panic' conversations and 'normal' conversations on social media platforms.
- Classification of Users in the marketing sectors into promoters or detractors based on their reviews on a product or brand is also an important application.
- Simpler use cases include filtering out spam and non spam emails.
- Sentiment Analysis for differentiating between positive,negative and neutral comments.
- Identification of genre or topics. This can be useful in recommendation systems where a document is assigned to a particular tag and other documents(videos/articles) are displayed to the user under the same tag.

III. Feasibility Study

- For successful classification of documents using NLP approaches, we must ensure that the quality of the training data and document classes is upto the mark. We must ensure that two categories must not overlap in terms of their content and must have lower degrees of similarities.
- For effective analysis , the content must be standardized and non relevant information must be removed.(preprocessing stage).
- Sometimes important keywords are not enough to classify the documents, the underlying context of the sentences is also essential. Classification can be difficult if the documents contain sarcastic comments where the meaning of the sentence is completely opposite.

IV. Current Systems

Most of the systems currently used for document classification use traditional approaches such as Naive Bayes, KNN and SVM. The problem with these traditional methods is that the models have low accuracy and have a high computational cost. The naive bayes performs poorly on unbalanced classes. The SVM algorithms for text classification are limited by the lack of transparency in results caused by a high number of dimensions and prone to overfitting. Although there are systems using neural networks like ANN and RNN, they are scarce.

V. Novelty/Innovativeness

Our project uses a novel approach of RNN which has a huge advantage over traditional systems like KNN, naive-bayes or SVM classifiers over speed, accuracy and various other factors.The innovativeness in our project lies in the use for RNN which is newer method used in Document classification in the domain of NLP.The methods currently used in the domain are mainly traditional methods which, though offer high accuracy, are computationally slow when it comes to complex conditions.We intend to evade this problem using RNN which are not only accurate but also fast.

VI. Literature Survey

This section specifies the various papers referred during the implementation of the project.

A. Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents

[1] This paper used Tf-Idf scores to classify documents into their specific domains. Calculation of Tf-Idf score of specific keywords of the text document with respect to each domain to classify the document into a particular category. The data consists of 20 different sites divided into 4 groups. Tf-IDf

Table 1. Domains & Websites

No.	Domains	Website's Count
1	.biz	5
2	.com	5
3	.edu	5
4	.org	5
Total	4	20

Fig. 1. Dataset used

provides relevance of keywords to some documents. Term Frequency (Tf) is used to measure how many times a term is present in the document. It is the measure of occurrence of numbers of the word divided by the total number of words. IDF lowers the importance of frequent words like 'of' , 'the' , 'a' etc. and gives more importance to infrequent words or words that give more meaning to the type of document. The formula for the same is ' ln(number of documents/no of documents containing the word)

The data was collected from 20 different sites grouped into 4 domains. Therefore each domain has 5 documents. The domains are .biz, .com, .edu and .org. The HTML,CSS tags are filtered out and the stop words are also removed from the documents. First the data is pre-processed , TF formula is applied on each word of each document. Then, the IDF formula is applied to each word.

After calculating TF, one has to check, if each word is found in every document or not and has to count the total number of documents in hand. Once these steps are completed, one can apply Inverse Document Frequency (IDF) formula to calculate IDF. For each document of each domain the term with the highest Tf Idf score is recorded. For test sentences, each word is taken and the domain corresponding to the highest tf-IDF score of the words is classified for the sentence. For an unseen document, the Tf Idf score of the words with respect to each domain is calculated and the domain which results in the highest TF-IDF score of the keywords is the class of the document.

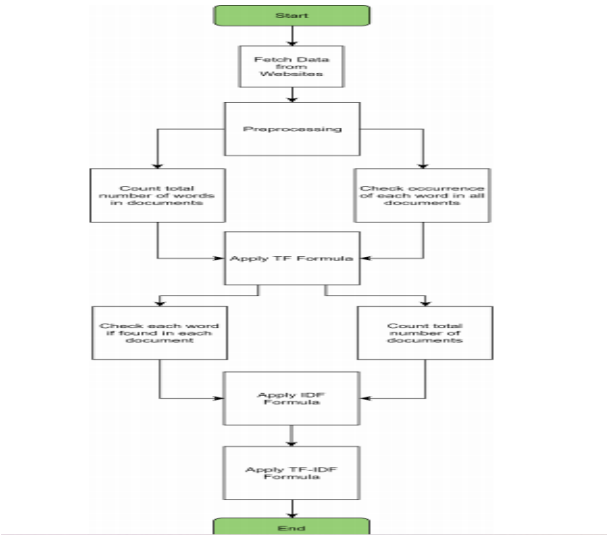

Fig. 2. Complete workflow

Merits and Demerits of the proposed method in the paper

Merits: The Tf-IDF scores give high importance to relevant keywords or important keywords of that particular document and hence can classify any document which belongs to any one of those 4 domains.

Demerits: Any change in the tense of a particular word cannot be handled by Tf-Idf. For example, 'run' and 'running' will be treated as two different words. However, this can be avoided by stemming or lemmatization in the preprocessing stage.Tf-Idf scores does not take the context of a particular keyword into consideration and fails to handle co-occurences of words. Calculating Tf-Idf scores of each keyword in a document with respect to each domain is a very tedious task when the number of documents per domain is large.

B. Document Classification using Neural Networks Based on Words

[2]Document classification is the key technique in text mining to organize the information in a supervised manner. Document classification is a task of classifying the documents into predefined categories. Neural network is a machine learning approach to classify the documents. In the proposed work, the keywords are extracted from the document and these are used for the classification purpose. This paper performs text categorization using back-propagation which is a technique for

ANN .
Data: created a database which contained a vocabulary list for each of the three categories i.e. the database had a table with three columns Physics, Chemistry and Biology, added words to each of the columns considering their relevance to that subject.
One of the most important developments in the neural networks is Back-Propagation learning algorithm. Backpropagation as an ANN is very useful in recognizing complex patterns and performing nontrivial mapping functions.Forward propagation through a neural network generates the output activations. Backward propagation through the training neural network pattern computes change in all the activations and hidden unit components. To update the parameters(weight and bias) multiply the delta or derivative of the parameters with the learning rate and subtract it from the original parameters. This is called Gradient Descent.
Stop words do not have so much meaning in a retrieval system and are a part of natural language. Stop-words should be removed from a text because texts look heavier and less important for analysis. For finding out the root/stem of a word, stemming technique is used.
To classify the documents, important keywords are extracted from the unclassified document and sent into the neural network. First important keywords are extracted from each paragraph of the document . These keywords are used to classify the paragraph they are part of . This is done for each paragraph of the document and the categories noted. The majority class is recorded as the category for that particular document. The keywords taken are Highest Frequency word, lowest frequency word and average frequency words from each paragraph.
For training of the neural network, 3 categories -Physics, chemistry and Biology were taken and each category was given a set of vocabulary or words relating to that particular topic. Each word was converted to a numeric vector before passing it to the network. Therefore each letter of the word is converted to it's ASCII format. For example : 'physicist' - [112 104 121 115 105 99 105 115 116].The neural network consists of 1 input layer(26 neurons) , 1 hidden layer(26 neurons) and 1 output layer(1 neuron). Every vocabulary list is normalized to length 26 by adding zeros to the vector. 1000 words for each category was passed into the neural network and the categories are encoded into numerical values.
Merits and Demerits of the proposed method in the paper
Merits: For the problems which cannot be solved sequentially or by sequential algorithms ANN provides the better solution. Back-Propagation is very useful in recognizing complex patterns and performing nontrivial mapping functions. Faster when compared to KNN as in KNN, it costs very much time for classifying objects if the number of training examples is large because it has to select a few objects by computing the distance of each test object using all the training examples. ANN's are preferred over Naive Bayes as features are dependent on each other in textual classification.
Demerits: Most frequent , least frequent and average frequent words are not enough to represent the paragraph . The importance of a word needs to be measured with respect to its occurrence in other documents as well.

C. Document Classification Using Artificial Neural Network[3]

Summary:This paper presents an artificial neural network approach which is hybrid of n-fold cross validation and training-validation-test approach for classification of data.

Datasets	Number of documents	Number of attributes (terms)	Number of classes (categories)
CNAE	1080	856	9
Db world (bodies) (stemmed)	64	3721	2
Db world (subjects) (stemmed)	64	229	2
Gender(Female)	3232	100	2
Gender(Male)	3232	100	2
Amazon	1500	50	50
Reuters 21578(Acq)	12897	100	2
Reuters21578(Earn)	12897	100	2
Reuters21578(Grain)	12897	100	2
Reuters21578(Corn)	12897	100	2
Reuters21578(Money)	12897	100	2

Fig. 3. Dataset used for classification

Methodology: The proposed methodology is based on n-fold training-validation-test approach which is combination of n-fold cross-validation with validation. In n-fold cross. After obtaining best validation performance the obtained ANN is employed for testing on test set. The proposed approach is hybrid approach of n-fold cross-validation and training-validation-test approach. Validation the dataset is initially divided into 'n' folds. Then during training each 'n' folds one by one acts as test set and remaining sets are treated as training set. Finally the mean error or accuracy obtained from each of 'n' test sets are the final result of classification.

Further in the proposed approach, the paper uses exhaustive weight initialization as it is observed that weight initialization in ANN affects the accuracy of results at the end so the paper selects the initial weight configuration ANN which gives best result at the end and hence the proposed method is called exhaustive validation and weight initialization.

Merits/Demerits: The advantage of the proposed approach is that it gives better results than the existing algorithms, as it has a capability to discover most optimum ANN for the given dataset.

The disadvantage being that the cost and time for building the ANN is huge when compared to traditional methods.

D. Text classification using artificial neural networks

Paper Summary: This paper describes text classification using neural networks, purpose of choosing the neural networks, advantages and the process of text classification. It also does a brief summary of major methods used.

Data: Various documents from the web.

Methodology and Approach: The data pre-processing is done in four steps: conversion, stemming, indexing and weighing.
In the initial step of document conversion, the document is changed over into plain content and the stop words like prepositions, participles are eliminated.
In stemming(second step), the affixes of the word are removed, and the root word is obtained.
The initial step of indexing includes the development of a super vector which holds all terms that show up in all documents of the corpus after that in the second step of indexing for dimensionality decrease is finished by separating a subset of the super vector dependent on certain rules that will work more effectively.
In the fourth step, term weighting is accomplished for each term and each document inside the corpus.

Merits/Demerits: The advantages of ANN are:
It can provide better solution for the problems which cannot be solved linearly or by using linear statistical classification techniques. Can learn even in the presence of noise. It does not force restrictions on the input variables like other prediction techniques. The disadvantage is that the cost of building is higher for more precise Neural networks and the parameter tuning is hard.

VII. Proposed Methodology

We propose to solve the problem statement by using Recurrent Neural Networks using LSTM for supervised classification of documents into their respective categories. The dataset must contain minimum 1000 entries per class for appropriate predictions. If we have 20 classes, then we 20,000 entries which needs to be split into training, testing and validation data. Before we pass the data into the Neural Network, we must preprocess the data and encode it in a form understandable by the algorithm. Preprocessing steps include removal of duplicate data, handling of imbalanced data (few classes being much more common than other classes), removal of stop-words (words that bring no meaning to the documents such as ‘a’,’an’,’the’ etc. These words can be removed using the nltk package.) , conversion of all Uppercase to lowercase letters, applying stemming or lemmatization to retain the root of the words.
Data - The dataset consists of a total of 2225 documents from the BBC news website spread across 5 topics namely:

- Business
- Tech
- Entertainment
- Politics
- Sports

In order to increase the volume of data, we extract each sentence from the document and assign the corresponding category to the same. The final dataset has 2688 entries for the business category, 2106 for entertainment,2684 for politics,2827 for tech,2585 for sports.
Visualization is done for better understanding of the dataset. The following image shows a word cloud for the Entertainment

category	business	entertainment	politics	sport	tech
count	2688.000000	2106.000000	2684.000000	2827.000000	2585.000000
mean	2427.925595	6761.707502	11130.289866	16170.277326	21119.177950
std	1405.891169	1104.590189	1425.700454	1481.937116	1377.112003
min	0.000000	4860.000000	8678.000000	13606.000000	18742.000000
25%	1210.500000	5801.500000	9894.750000	14895.000000	19927.000000
50%	2429.000000	6753.000000	11121.000000	16171.000000	21120.000000
75%	3648.500000	7729.500000	12369.500000	17455.000000	22304.000000
max	4859.000000	8677.000000	13605.000000	18741.000000	23519.000000

Fig. 4. Dataset Description

category. Larger the word, more prominent and more important it is to the category. Once we complete all the cleaning and the

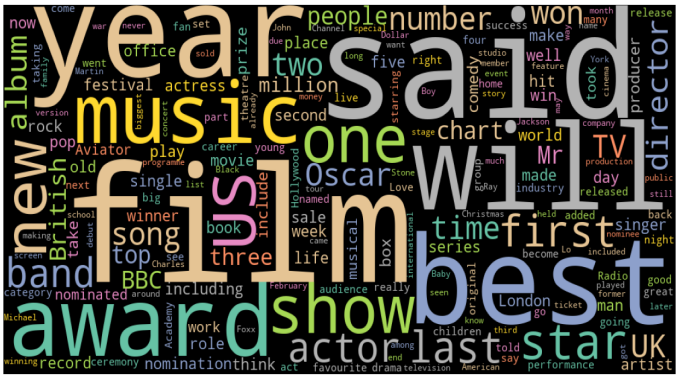

Fig. 5. Word Cloud for Entertainment

visualization, we split the data into 80 percent training and 20 percent testing. We use the inbuilt module ’Tokenizer’ from keras to convert the sentences to sequences. Each vector must be of same size, therefore we are required to pad zeroes to end of each vector. The labels are one hot encoded. Following is the architecture of our network.

```
Model: "sequential_10"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_10 (Embedding)     (None, None, 8)           8000
_________________________________________________________________
lstm_10 (LSTM)               (None, 8)                 544
_________________________________________________________________
dense_18 (Dense)             (None, 8)                 72
_________________________________________________________________
dense_19 (Dense)             (None, 6)                 54
=================================================================
Total params: 8,670
Trainable params: 8,670
Non-trainable params: 0
_________________________________________________________________
```

Fig. 6. Neural Network Architecture

VIII. Results

Our model was able to achieve 82.93 percent validation accuracy and 90.88 percent training accuracy. The loss func-

tion used was categorical crossentropy with 'Adam' as the optimizer. We ran the algorithm for 15 epochs.With the results achieved, we can conclude that our model is slightly overfitting. Natural Language Applications are prone to overfitting due to a large number of unseen vocabulary. The vocabulary size plays a huge role in the generating the results. Another way that is possible to decrease overfitting is to use Tf-Idf vectorization to generate sequences. Tf-Idf prioritzes rare words and discards common words. We may also use regularization techniques in our architecture to prevent overfitting.

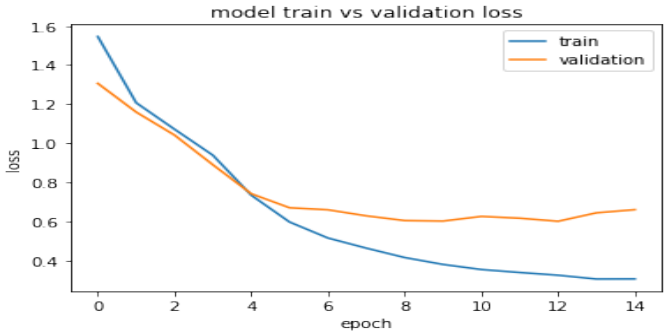

Fig. 7. Training loss vs Validation loss

IX. References

[1] Qaiser, Shahzad Ali, Ramsha. (2018). Text Mining Use of TF-IDF to Examine the Relevance of Words to Documents. International Journal of Computer Applications. 181. 10.5120/ijca2018917395.

[2]Document Classification using Neural Networks Based on Words

[3]Tripathi, Kshitij Vyas, Rajendra Gupta, Anil. (2019). Deep Learning through Convolutional Neural Networks for Classification of Image: A Novel Approach Using Hyper Filter. International Journal of Computer Sciences and Engineering. 7. 10.26438/ijcse/v7i6.164168

[4]Prasanna, P. Rao, Dr. (2018). Text classification using artificial neural networks. International Journal of Engineering and Technology(UAE). 7. 603-606. 10.14419/ijet.v7i1.1.10785.