International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 06 | June-2018 www.irjet.net p-ISSN: 2395-0072

Automated Document Summarization and Classification Using Deep Learning

Krushna Sharma[1], Avinash Gaikwad[2], Swapnil Patil[3], Pradeep Kumar [4], D.P. Salapurkar[5]

[1,2,3,4] *B.E. (Computer Engineering), Sinhgad College of Engineering, Pune, Maharashtra, India*
[5]*Assistant Professor, Dept. of Computer Engineering, Sinhgad College of Engineering, Pune, Maharashtra, India*

Abstract – *The exponential growth of the Internet has led to great deal of interest in developing useful and efficient tools and software to assist users in searching the web for relevant documents. Document classification is generally defined as content-based assignment of one or more predefined categories to documents. Document classification appears in many applications, including email-filtering, news monitoring, etc. It is not feasible to classify these documents manually and present automated classification methods have drawbacks like low accuracy and dependency on humans for document tagging.*

The proposed system uses deep learning methods to speed up the classification process and recommend relevant documents. The proposed deep learning algorithm -'Recurrent Neural Network with Convolutional Neural Network' helps in construction of a robust classifier model using variety of data for training. This classifier can then be improvised to classify documents in a business database automatically.

***Key Words*: Summarization, Classification, Neural Network, Deep Learning, Recurrent Neural Network (RNN), Convolutional Neural Network (CNN), Recurrent Convolutional Neural Network (RCNN).**

1. INTRODUCTION

The automated categorization (or classification) of texts into predefined categories has witnessed a booming interest in the last 10 years, due to the increased availability of documents in digital form and the ensuing need to organize them. In the research community the dominant approach to this problem is based on machine learning techniques: a general inductive process automatically builds a classifier by learning, from a set of pre-classified documents, the characteristics of the categories. The advantages of this approach over the knowledge engineering approach (consisting in the manual definition of a classifier by domain experts) are a very good effectiveness, considerable savings in terms of expert labor power, and straightforward portability to different domains.

Until the late '80s the most popular approach to TC, at least in the "operational" (i.e., real world applications) community, was a knowledge engineering (KE) one, consisting in manually defining a set of rules encoding expert knowledge on how to classify documents under the given categories. In the '90s this approach has increasingly lost popularity (especially in the research community) in favor of the machine learning (ML) paradigm, according to which a general inductive process automatically builds an automatic text classifier by learning, from a set of pre-classified documents, the characteristics of the categories of interest. The advantages of this approach are an accuracy comparable to that achieved by human experts, and a considerable savings in terms of expert labor power, since no intervention from either knowledge engineers or domain experts is needed for the construction of the classifier or for its porting to a different set of categories.

The proposed system implements Recurrent Neural network along with Convolutional neural network to build the classifier model. Only summary of document is used for classification phase which speeds up the training phase considerably.

1.1 Background and Basics

With the dramatic growth of the Internet, people are overwhelmed by the tremendous amount of online information and documents. This expanding availability of documents has demanded exhaustive research in the area of automatic text summarization. A summary is defined as "a text that is produced from one or more texts, that conveys important information in the original text(s), and that is no longer than half of the original text(s) and usually, significantly less than that". Automatic text summarization is the task of producing a concise and fluent summary while preserving key information content and overall meaning. In recent years, numerous approaches have been developed for automatic text summarization and applied widely in various domains. For example, search engines generate snippets as the previews of the documents. Other examples include news websites which produce condensed descriptions of news topics usually as headlines to facilitate browsing or knowledge extractive approaches. Automatic text summarization gained attraction as early as the 1950s. An important research of these days was for summarizing scientific documents.

Document classification or document categorization is a problem in library science, information science and computer science. The task is to assign a document to one or more classes or categories. This may be done "manually" (or "intellectually") or algorithmically. The intellectual classification of documents has mostly been the province of library science, while the algorithmic classification of documents is mainly in information science and computer

science. The problems are overlapping, however, and there is therefore interdisciplinary research on document classification. In recent years, deep learning algorithms have drastically improved the speed and efficiency of automated document classification.

2. SURVEYED RESEARCH

Saif alZahir and Qandeel Fatima in [6] suggests

- Graph based summarization is simple to implement and does not rely on the calculations of the cosine similarity between sentences to rank them in the summary

Guibin Chen, Deheng Ye and Zhenchang Xing in [1] suggests:

- A convolutional neural network (CNN) and recurrent neural network (RNN) based method is capable of efficiently representing textual features and modeling high-order label correlation with a reasonable computational complexity
- The power of RCNN is affected by the size of the training dataset
- Larger the dataset better is the accuracy
- Smaller datasets may result in over fitting

Md. Saiful Islam in [2] suggests:

- A method which depends on TF-IDF - a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus
- TF-IDF is based on the bag-of-words (BOW) model, therefore it does not capture position in text, semantics, co-occurrences in different documents, etc

Sebastian Suarez Benjumea in [5] suggests:

- Relevance extraction and more non-overlapped clustering
- This process is more time and compute intensive

3. PROPOSED SYSTEM

3.1 Data Pre-Processing

This process removes all the special characters and irrelevant information such as images and diagrams from the given document. Here, the textual data is also processed to filter out special symbols, stop-words, etc. Preprocessing ensures that clean and relevant data is provided for summarization and classification modules.

3.1 Text Summarization

The main idea of summarization is to find a subset of data which contains the "information" of the entire set.

The proposed system uses Graph Based Text Summarization method which is encouraged by Google's Page Rank algorithm. Here an intersection matrix is generated based on common words between sentences and score is calculated for each sentence. Based on the score, some percent (20) of top sentences are selected to form summary.

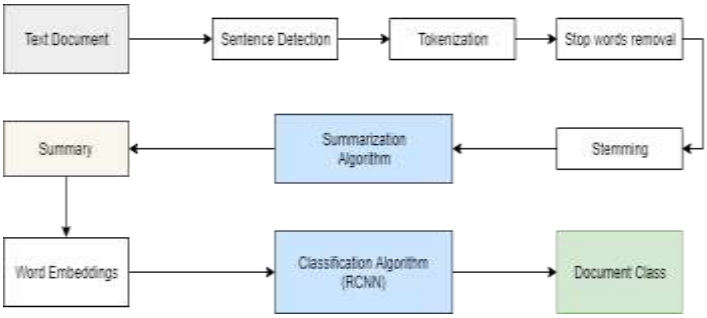


Fig -1: Architecture Diagram

3.3 Text Classification

Text classification is an algorithm which assigns one of the predefined classes to a document according to the type of content it holds. This is done by using Natural language processing along with machine learning techniques such as SVM (Support Vector Machine), Deep learning models (CNN, RNN, RCNN).

Supervised learning methods are efficient when it comes to classification of text documents providing better results with far better classification accuracy than unsupervised algorithms. Text classification has multiple applications such as spam detection, web content analysis, business applications (Product trend classification according to the feedback given by users), Analytics; Organizing documents in the ETL warehouse according to their contents so the analyst could only focus his time on specific domains saving his time.

Word2Vec model provides contextual representation of the document that preserves the context of sentences. Word2Vec model is developed by using unsupervised method such that words relevant to each other are clustered together.

The model provides a numerical representation from textual representation. Now this representation can be fed to deep learning methods.

4. IMPLEMENTATION

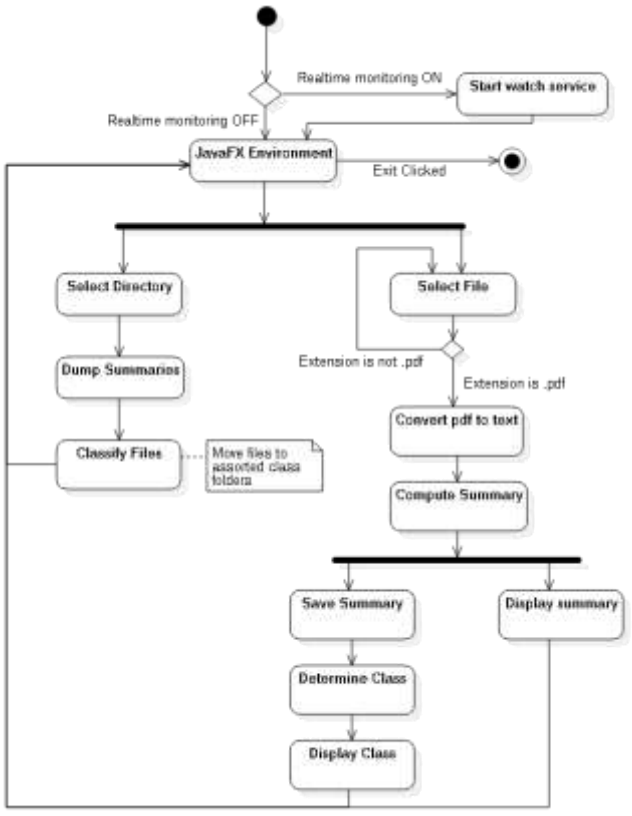


Fig -2: Activity Diagram

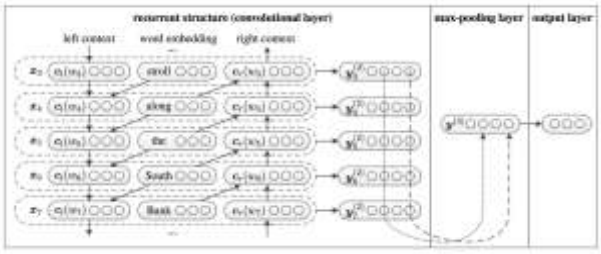
4.1 Graph Based Summarization Algorithm

This method is unique in that it produces a multi-edge-irregular-graph that represents words occurrence in the sentences of the target text. This graph is then converted into a symmetric matrix from which we can produce the ranking of sentences and hence obtain the summarized text using a threshold. To test our method performance, we compared our results with those from the most popular publicly available text summarization software using a corpus of 1000 samples from 5 different applications: literature, politics, religion, science and sports. The simulation results show that the proposed method produced better or comparable summaries in all cases. The steps include:

1. Extract sentences from given raw text
2. Group sentences in paragraphs
3. Create intersection matrix
4. Calculate no of common words between two given sentences and then compute sentence score as: Score = No. of common words / Total no. of words /2
5. Sort sentences based on score
6. Display top 20 percent sentences as summary

The proposed method is fast and can be implement for real time summarization.

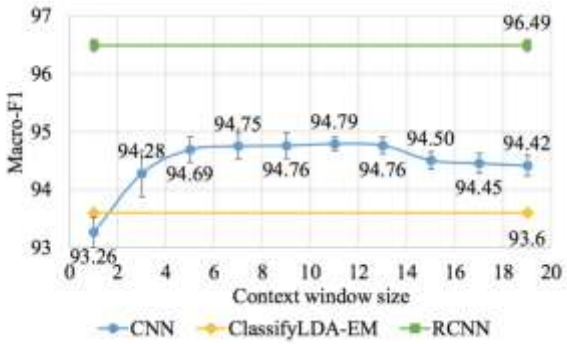
4.2 Recurrent Convolutional Neural Network (RCNN)



Recurrent Neural networks capture contextual information by maintaining a state of all previous inputs. The problem with RNNs is that they're a biased and favor more recent inputs. CNNs can learn important words or phrases through selection via a max pooling layer. However, processing text is difficult with CNNs because learning an optimal kernel size is challenging.

These obstacles could be avoided by implementing an algorithm that will combine their effectiveness along with the disadvantages mentioned. Such an algorithm is called RCNN (Recurrent Convolutional Neural Network).

While preserving the state of previous input by employing a recurrent (long-short term memory layer) with convolutions to remove the bias for previous inputs. Results obtained from this algorithm are provided below.



Model used in this system provides accuracy about 97.76%, trained on data from various domains.

5. CONCLUSIONS

The Page-Rank inspired algorithm used for text summarization gives better results than other algorithms due to accuracy and more abstractive like results. Classification algorithm (i.e. deep learning methods) propose better semantic information and context preservation. Due to summarization the classification becomes more intuitive and

simple. Classifier model can be trained to incorporate classes corresponding to any business requirement provided that sufficient training data is available. Accuracy of classifier can be increased by increasing the variety of documents in dataset. Convolutional and Recurrent Neural Network (RCNN) algorithm ensures that semantic correlations are also considered during classification

REFERENCES

[1] Guibin Chen, Deheng Ye, Zhenchang Xing - "Ensemble Application of Convolutional and Recurrent Neural Networks for Multi-label Text Categorization", IEEE 2017

[2] Md. Saiful Islam, "A Support Vector Machine mixed with TF-IDF Algorithm ti Categorize Document", IEEE 2017

[3] Sumya Akter, "An Extractive Text Summarization Technique for Bengali Document(s) using K-means Clustering Algorithm" , IEEE 2017

[4] Rui Zhao and Kezhi Mao, "Fuzzy Bag-of-Words Model for Document Representation", IEEE 2016

[5] Sebastian Suarez Benjumea, "Genetic Clustering Algorithm for Extractive Text Summarization", IEEE 2015

[6] Saif alZahir and Qandeel Fatima, "New Graph-Based Text Summarization Method", IEEE 2015