International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 04, April-2022

Novelty Detection in Text Document using Convolutional Neural Network

Shalini Ranjan[1], Supriya P. Panda[2]
[1]M. Tech ,(pursuing in Computer Engineering and Networking), Department of Computer Science & Engineering, FET, MRIIRS, Surajkund, Faridabad (Haryana)
[2]Department of Computer Science & Engineering, FET, MRIIRS, Surajkund, Faridabad (Haryana)

Abstract:- **The increasing growth in the amount of information and number of users leads to difficulty in providing novel search services for users. Due to a large number of documents on the web, analysis of every collection of document is not feasible or possible. Thus, there is a need of Novelty detection activity. The web is a Search Engine, Search engine is a program that searches for documents according to the keywords provided by user and returns a list of the documents where the keywords are matched. Due to the increasing information over net it is necessary to retrieve the relevant and novel information according to the query user's query. Hence, it makes it essential to find means for providing relevant and novel information from the query given by the user. Therefore, Novelty detection comes into the picture. The main purpose of novelty detection is to provide with user a list of documents that are relevant to the query given and also contain new information according to the user's information request. In this work, Convolutional Neural Network (CNN) based model is proposed for novelty detection, which will provide relevant and novel documents according to the query given by the user. This system extracted data from the web and used that as a dataset. After collecting Uniform Resource Locator(URL) from the web those URLs are stored in the file and from them the novel document is selected based on Relative Document Vector(RDV). The proposed system concludes with feature representations from a target document relating to the source documents using a CNN.**

Key words: Convolutional Neural Network(CNN), Novelty detection, Search engine.

1. INTRODUCTION

1.1 Information Retrieval is referred to as a process that responds to a user query by taking into consideration a group of documents and returning a document list in a sorted manner that provides all the relevant information about the queries to the user.

Information retrieval (IR) is the process of getting all the data that is relevant to the information needed. When given a query, the system gives results related to any word present in the Query [5]. Information retrieval also searches the information in a document, checks the data that describes the data, and also for texts, images of sounds.

An information retrieval is a system that is used to store the information that need to be processed, searched, and retrieved according to user query. Information retrieval is also called as document retrieval. The result of a document retrieval system is a list of documents that are ranked by the relevance scores that are calculated by the system. The system assumes that a document with a higher relevance score is more relevant to the user's query than other document. The purpose of novelty detection is to provide a user with a list of document that are relevant and contain new information. The goal of novelty detection is to get useful information without reading all documents, which is usually time-taking task. Novelty detection is a step forward in document retrieval [6].

1.2 Text Classification

Text classification is the method of assigning/giving tags or categories to text according to its provided content. It is one of the basic tasks in Natural Language Processing (NLP) such applications like sentiment analysis, topic labelling, spam detection, email filtering etc. are some of the examples where text classification is used [3].

1.3 Novelty Detection

Novelty detection is the distinguishing proof of new information or sign that an AI framework doesn't know about during preparing [4]. Oddity recognition is one of the major prerequisites of a decent arrangement or distinguishing proof framework since in some cases the test information contains data about articles that were not known at the hour of preparing the model. In this paper we give best in class audit in the zone of oddity identification dependent on factual methodologies [1].

1.4 Text Classification Using Convolutional Neural Network

The first thing needed in Convolutional Neural Network is a way to transform a piece of text into a vector of numbers after that CNN can be applicable with them. One of the most typical tasks in Natural language processing (NLP) where CNN are used is text classification, i.e., classifying a text into a set of pre-determined categories by considering n-grams means each and every text in our dataset is represented thousands of dimensions in a vector form, each one representing the count of one of the words of the text, this is what we called as Word embedding.

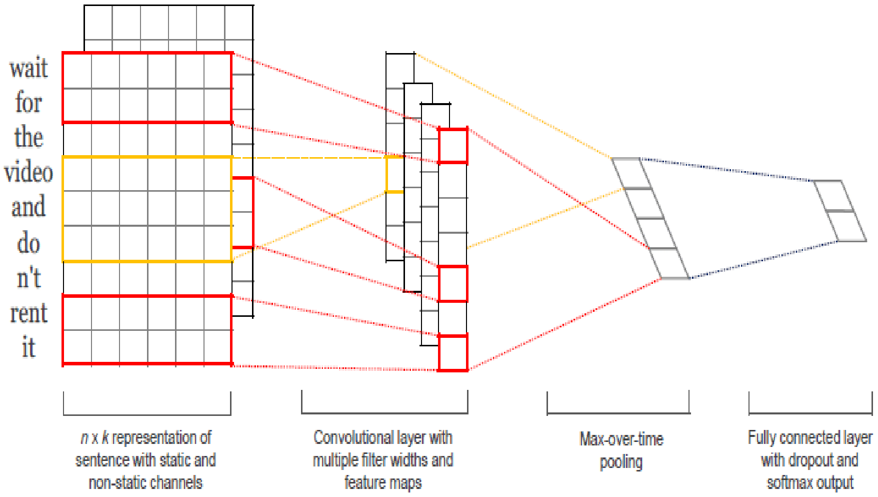

wait
for
the
video
and
do
n't
rent
it
n x k representation of sentence with static and non-static channels
Convolutional layer with multiple filter widths and feature maps
Max-over-time pooling
Fully connected layer with dropout and softmax output

Fig. 1: CNN Architecture model for text classification.

There have been many efforts for enhancing the CNN-model architectures [9] as shown in Fig.1. Liu et al. proposed a new CNN model that makes two modifications to the architecture used by Kim. Firstly, a dynamic max-pooling scheme is taken into account to captures more fine-grained features from different regions of the document. Second, a hidden bottleneck layer is used between the pooling and output layer to learn compact document representations to reduce the size and improve the model performance.

2. LITERATURE REVIEW

Sushil Kumar and Komal Kumar Bhatia, (2019) proposed a cluster based approach for novelty detection which will provide the novel and relevant documents for the information need. Based on the user query the incoming stream of documents will be clustered using K- means algorithm. The novel document based on the query and filter out the redundant document.

D.Miljkovic, (2010) states about the review of novelty detection. Novelty detection methods try to identify outliers that differs from the distribution of ordinary data. There are different approaches applied i.e., Classification based approach, Nearest Neighbour, Clustering based approach and others.

X. Zang, J. Zhao, and Y. Lecun, (2015), offers an empirical exploration on the use of character-level convolutional networks (ConvNets) for text classification. Several largescale datasets are constructed to show that character-level convolutional networks could achieve state-of-the-art or competitive results.

Tirthankar Ghosa, Vignesh Edithal (2018) aimed at automatically classifying an incoming document as novel or non-novel on the basis of documents already seen by the system. The rapid growth of documents across the web has necessitated finding means of discarding redundant documents and retaining novel ones. Capturing redundancy is challenging as it may involve investigating at a deep semantic level. Techniques for detecting such semantic redundancy at the document level are scarce. In this work we propose a deep CNN based model to classify a document as novel or redundant with respect to a set of relevant documents already seen by the system.

H. T. Le, C. Cerisara, and A. Denis, (2018) shows whether the convolutional networks need to be deep for text classification or not, here, display the importance of depth in convolutional models for text classification, either when character or word inputs are considered.

3. PROBLEM IDENTIFICATION

The relevancy and novelty of documents are very important for the user. But Search Engine provides a list of documents that can be relevant or redundant. Because of that, users require more time to search for the document they need. Convolutional Neural Network helps in filtering the novel and relevant documents according to the query given by the user. Therefore, the users have to put their minimum effort to find novel and relevant documents according to their query. Search engine filter results, the information according to the query given by the user, which may provide relevant results to the users. The most common problem is the number of repetitive title tags used. This is where Google faced issues and something that is easily fixed. Therefore, every page on the website has a unique tag.

3.1 Research Objectives

The research objectives aim:

i. To understand and identify whether an incoming document is a novel or non-novel based on prior known documents and as per the users' query.

ii. To apply Web Scrapping to extract data.

iii. To analyse that the extracted data is stored in data.csv file after extracting the data from the web. After this data will be pre-processed. Here, Tokenization comes into picture it is a way of turning an important bit of information, then after removing stop words from document list are stored in csv file.

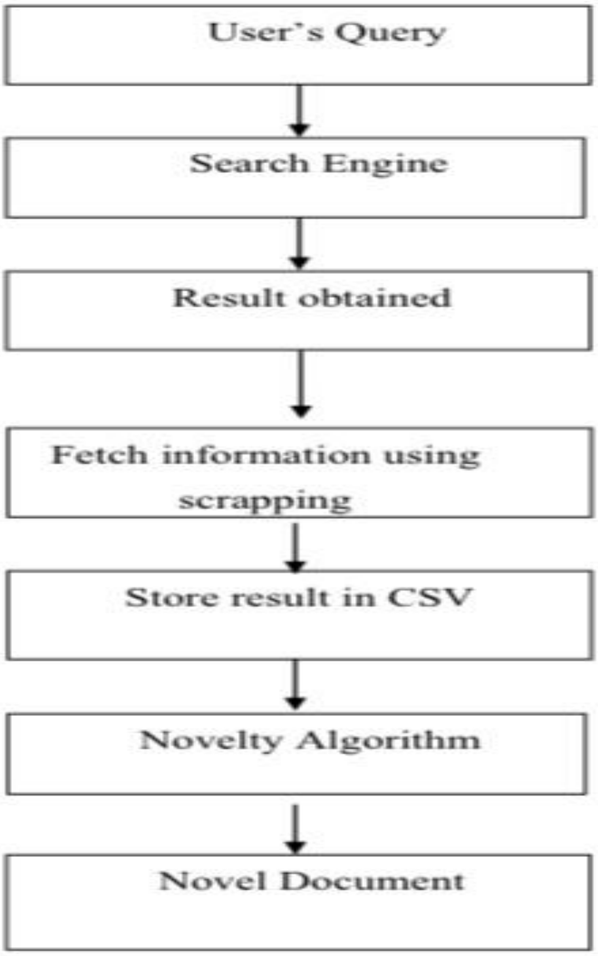
iv. To design the model, data is split into two sub set i.e. training and testing, now data is train and test by using CNN. Here 1D convolution is used, filters are applied. K- Nearest Neighbors (KNN) that is import by using libraries that are predefined in python and then comparison is done in Convolutional Neural Network on the same.

4. RESEARCH METHODOLOGY

4.1 This research will be carried out in the following steps to achieve the objectives of the research:

Step1: Extract data from the web.
Step 2: Data pre-processing
Step 3: Calculate Term Frequency and Inverse Term Frequency
Step 4: Extract Relative Document Vector
Step 5: Extract Feature
Step 6: Apply CNN to find the novel result.
Step 7: Final Result. As shown in Fig. 2&3

Fig.2: General Architecture for Novelty Detection



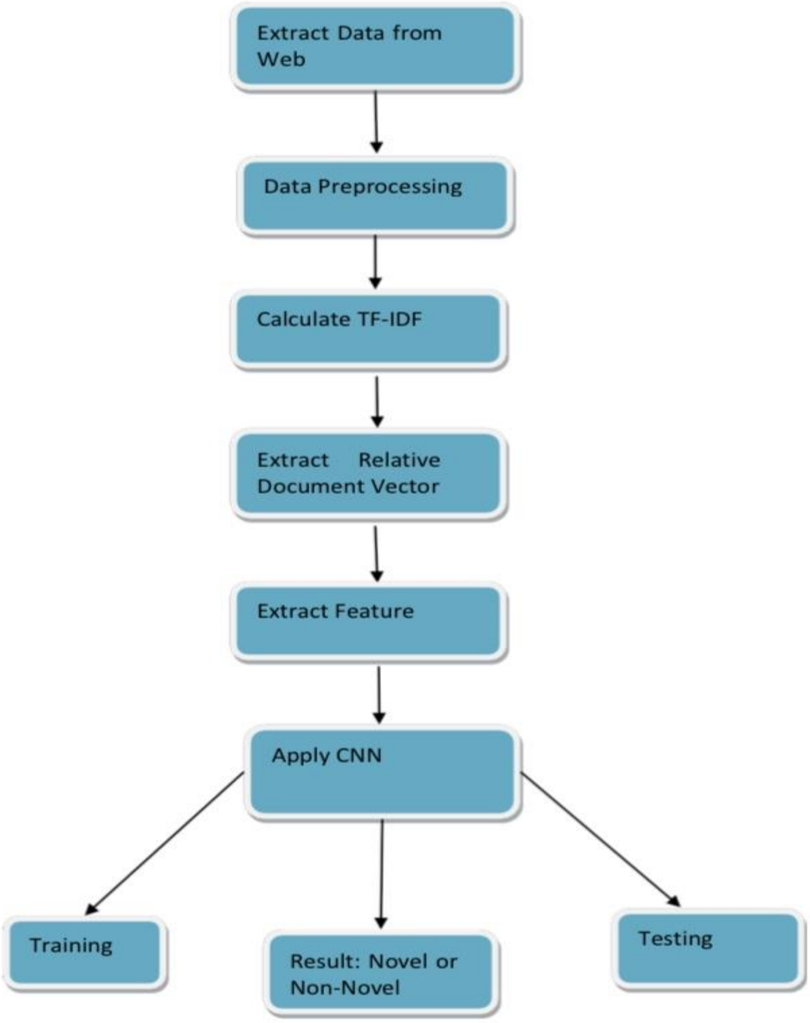


Fig.3: Proposed Architecture

In this proposed system, extracting data from the web is used as the dataset. The proposed system learns feature representations from a target document by considering the source documents using a CNN. The Relative Document Vector (RDV) is applied to the data extracted from the web [8]. The general architecture of Novelty Detection is shown in Fig. 3.

4.2 WEB SCRAPING

Web scraping is a method that is used to extract large amounts of data from websites by just writing code. The data on the websites are not structured. Web scraping helps to collect the unstructured data and store this data in a structured form [10].

4.2.1 Data Scraping from a website:

When the code is written for web scraping and run that code then a request is sent to the URL as mentioned/given by the user. The code then, parses the pages using an lxml parser, finds the data according to the request given, and extracts it.

Some steps that are followed to extract data using web scraping with python

i. Search the URL, you want to scrape.
ii. Look at the Page carefully.
iii. Search the data, you want to extract.
iv. Write the web scraping code.
v. Run that code and extract the data from the given URL.
vi. Store the extracted data in the format required such as .csv, .txt ,etc.

5. RESULTS & DISCUSSION

5.1 Software and Hardware Requirements

The minimum software and hardware requirements for implementing the proposed work. The implementation of algorithm is going to perform on Processor: intel core 3, running under windows 10 operating system. The algorithm will be done under Python software. The implementation will include Anaconda Notebook Python is a free open source distribution of the Python language.

5.2 The various step for implementation is shown below:

1. Data crawling from the web using Web Scrapping.
2. Crawled data is stored in data.csv file after crawling the data from web.
3. Data will be pre-processed.
4. Then, data will be split into two sub set i.e. training and testing using CNN.
5. Data is going to be train using KNN and then there will be comparison between CNN and KNN.

6. CONCLUSION AND FUTURE SCOPE

In this work, a comparison will be done between CNN and KNN for document- level novelty detection to distinguish a document as novel or redundant concerning a set of relevant documents. The proposed system learns feature representations from a target document to the source documents using a Convolutional Neural Network (CNN). Relative Document Vector (RDV) is applied to extract data.

6.2 Future work can be enhanced version of the proposed architecture can be extended to improve the efficiency of novelty detection methods. Many other techniques such as semantic similarity and text summarisation can be used to increase efficiency. In the existing paper, a clustering- based approach for novelty detection has been investigated. The incoming stream of documents based on the query has clustered using K- means algorithm and then cluster heads are calculated. This paper uses KNN [6]. The proposed architecture aims to increase efficiency by using CNN.

REFERENCES

[1] Li, Xiaoyan, Croft, W. Novelty detection based on sentence level patterns, pp. 744-751, 2006.

[2] H. T. Le, C. Cerisara, and A. Denis, “Do convolutional networks need to be deep for text classification?”, in Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[3] Hicham El Boukkouri, Text Classification: The First Step towards NLP Mastery, 2020.

[4] Novelty Detection: A perspective from Natural Language Processing, Tirthankar Ghosal, Tanik Saikh, Tameesh Biswas, Asif Ekbal, Pushpak Bhattacharyya, pp. 79-84, 2022.

[5] Parul Kalra Bhatia, Tanya Mathur, Tanya Gupta. Survey paper on Information Retrieval algorithm and personalised information retrieval concept, pp. 14-17, 2013.

[6] Sushil Kumar and Komal Kumar Bhatia,” Clustering based approach for novelty detection in text documents”, in Asian Journal of Computer Science and Technology, ISSN:2249-0701 Vol.8 No. 2, pp.116-121, 2019.

[7] Tirthankar Ghosal, Vignesh Edithal, Asif Ekbal, Pushpak Bhattacharyya. “A Deep Neural Solution to Document-level Novelty Detection, “Proceedings of the 27th International Conference on Computational Linguistics, pp. 2802-2813 Santa Fe, New Mexico, USA, 2018.

[8] X. Zang, J. Zhao, and Y. Lecun. “Character-level convolutional networks for text classification”, in Advances in neural information processing systems, pp. 649–657, 2015.

[9] Y. Kim, “Convolutional neural networks for sentence classification”, in EMNLP Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, pp. 1746-1751, 2014.

[10] https://www.wikipedia.org/