IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

News Classification Using CNN

Shaik Shah Ali Safwan

Co- Authors: -Ashish Pandey, Shaik Mohammad Tausif Jabbar

Brindavan Institute of Technology and Science, Kurnool

ABSTRACT:

As Internet usage is growing drastically there are tons of web content getting generated every day on social media, websites, etc. Over 90% of the data that is generated is unstructured, to make them structured like to make them categorize, summarize them, making some conclusions we must read the entire data which is very huge. So, to overcome this problem this paper proposes a new text classification method using Natural Language Processing (NLP) techniques and Deep Learning algorithms. This paper has customized Deep Learning algorithms such as CNN and Transformers, and Transformers + CNN for classifying the news text. To check the performance of the classification model Accuracy metric has been used and found the best model which was integrated with the front-end

KEYWORDS:

News, Deep learning, CNN, RNN, LSTM, Classification, NLP, Algorithm, Transformers.

INTRODUCTION:

With the explosion of news sources in today's digital age, ranging from online media and social media to traditional outlets like TV, newspapers, and radio, the need for efficient categorization and organization of this information has become paramount. Traditionally, this process relies on manual review and classification, often by dedicated teams, which can be time-consuming, laborious, and prone to human error. Text classification, a fundamental task within Natural Language Processing (NLP), offers a compelling solution to this challenge. It employs machine learning algorithms to automatically categorize textual content into predefined categories, such as "Politics", "Crime", "Sports", "Business", "Health", etc. This automated approach offers several key advantages:

- Efficiency: Text classification algorithms can process massive amounts of text data significantly faster than human reviewers, enabling near real-time categorization and significantly reducing processing time.
- Accuracy: By leveraging large datasets of labelled text and sophisticated machine learning techniques, text classification models can achieve high accuracy levels, often surpassing human performance in consistency and objectivity.
- Scalability: These algorithms can be easily scaled to handle continuously growing news volumes, ensuring efficient and consistent categorization even as the information landscape evolves.
- Cost-effectiveness: Automating the categorization process significantly reduces the manpower requirements compared to manual approaches, leading to substantial cost savings.
- Improved insights: By automatically identifying the topics and themes within news articles, text classification facilitates deeper analysis and understanding of current events and trends, enabling more informed decision-making.

The process of text classification involves two key steps:

1. Data Preprocessing: This stage involves cleaning and preparing the text data for analysis, including removing stop words, stemming or lemmatization, and tokenization.
2. Feature Extraction: The text is then transformed into numerical features that can be used by machine learning algorithms. This can involve bag-of-words representation, n-grams, and other techniques.

Once the features are extracted, various machine learning models can be trained to perform the actual classification task. Popular models include Naive Bayes, Support Vector Machines (SVM), and deep learning architectures like Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs).

OBJECTIVES:

The following are the objectives of this project

- This project is aimed to build automatic text classifiers to categorize the text which saves and helps the user time.
- It uses Natural Language Processing (NLP) to classify the text and gives the predefined class labels as output
- Creating Web UI for taking the text and displaying the predicted class label as output or improve the existing functions on each new software launch. Agile methodologies allow designing software to break down into manageable parts called "user stories." Pedrycz suggests that this underline Agile 's value for the consumer, allowing developers to deliver quicker input loops and maintain product compatibility with business needs [3]. Agile also supports adaptive planning, evolving development, early and ongoing delivery, and constant improvement to enable developers to respond quickly and flexibly to customer needs, software, and other external factors.

SYSTEM SPECIFICATIONS:

To be used efficiently, all computer software needs certain hardware components or other software resources to be present on a computer. These prerequisites are known as (computer) system requirements and are often used as a guideline as opposed to an absolute rule. Most software defines two sets of system requirements: minimum and recommended. With increasing demand for higher processing power and resources in newer versions of software, system requirements tend to increase over time. Industry analysts suggest that this trend plays a bigger part in driving upgrades to existing computer systems than technological advancements. Non-Functional Requirements Non-functional requirements are the functions offered by the system. It includes time constraints on the development process and standards. The non-functional requirements are as follows:

- Speed: The system should process the given input into output within appropriate time.
- Ease of use: The software should be user friendly. Then the customers can uneasily, so it does not require much training time.
- Reliability: The rate of failure should be less then only the system is more reliable
- Portability: It should be easy to implement in any system.

The specific requirements are:

- User Interfaces: The external users are clients. All the clients can use this software for indexing and searching.
- Hardware Interfaces: The external hardware interface used for indexing and searching is personal computers of the clients. The PC's may be laptops with wireless LAN as the internet connections provided will be wireless
- Software Interfaces: The operating Systems can be any versions of windows.
- Performance Interfaces: The PC's used must be at least Pentium 4 machines so that they can give optimum performance of the products.

SOFTWARE SPECIFICATIONS:

Software specification deal with defining software resources requirements and prerequisites that need to be installed on a computer to provide optimal functioning of an application. These requirements or prerequisites are generally not included in the software installation package and need to be installed separately before the software is installed. Software Requirements:

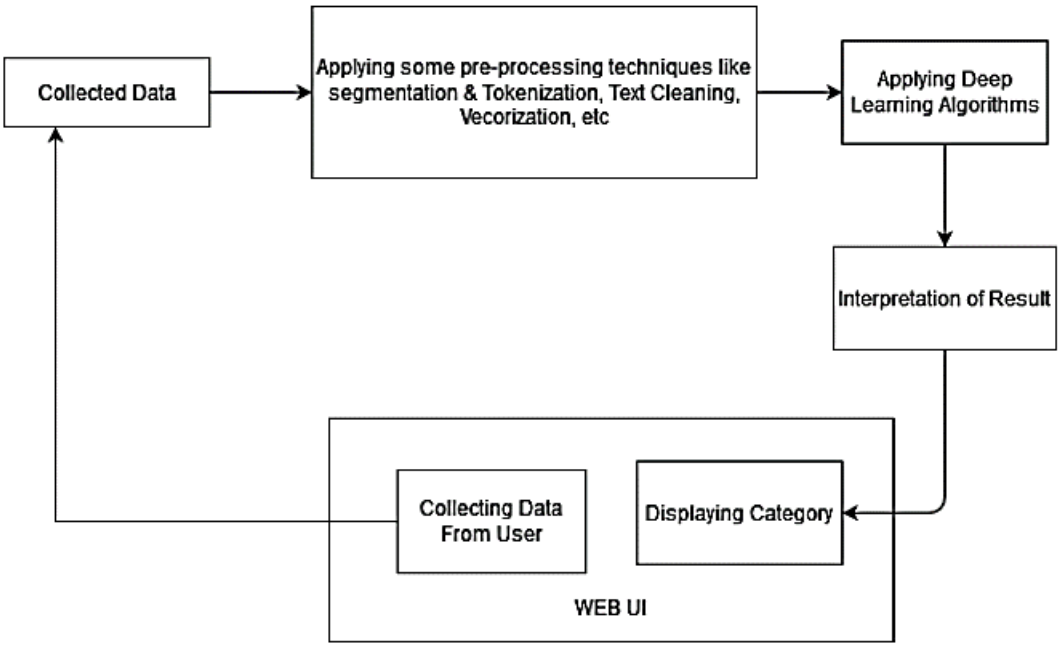
- Operating System : Windows 10
- Platforms : Jupyter Notebook, Anaconda
- Language : Python

HARDWARE SPECIFICATIONS:

The most common set of requirements defined by any operating system or software application the physical computer resources, also known as hardware, A hardware requirements list is often accomplished by a hardware compatibility list, especially in case of operating systems. An HCL lists tested, compatible, and sometimes incompatible hardware devices for a particular operating system or application. All computer operating systems are designed for computer architecture. Most software applications are limited to operating systems running on architectures. Although architecture-independent operating systems and applications exist, most need to be recompiled to run on a new architecture. The power of the central processing unit (CPU) is a fundamental System requirement for any software. Most software running on x86 architecture define processing power as the model land the clock speed of the CPU. Many other features of a CPU that influence its speed and power, like bus speed, cache, and MIPS are often ignored. This definition of power is often erroneous, as AMD Athlon and Intel Pentium CPUs at similar clock speed often have different throughput speeds

DESIGN & IMPLEMENTAION:

In deep learning, a convolutional neural network (CNN, or ConvNet) is a class of artificial neural network (ANN), most applied to analyse visual imagery CNNs are also known as Shift Invariant or Space Invariant Artificial Neural Networks (SIANN),
based on the shared weight architecture of the convolution kernels or filters that slide along input features and provide translation-equivariant responses known as feature maps. Counter-intuitively, most convolutional neural networks are not invariant to translation.



Due to the down sampling operation, they have applications in image and video recognition, recommender systems, image classification, image segmentation, medical image analysis, natural language processing, brain–computer interfaces, and financial time series. A transformer is a deep learning model that adopts the mechanism of self-attention, differentially weighting the significance of each part of the input data. It is used primarily in the fields of natural language processing (NLP) and computer vision (CV). Like recurrent neural networks (RNNs), transformers are designed to process sequential input data, such as natural language, with applications towards tasks such as translation and text summarization. However, unlike RNNs, transformers process the entire input all at once. The attention mechanism provides context for any position in the input sequence. For example, if the input data is a natural language sentence, the transformer does not have to process one word at a time. This allows for more parallelization than RNNs and therefore reduces training times. Transformers were introduced in 2017 by a team at Google Brain and are increasingly the model of choice for NLP problems, replacing RNN models such as long short-term memory (LSTM). The additional training parallelization allows training on larger datasets

WORKING METHODOLOGY:

This system has two sections hardware and software. The hardware consists of a laptop or mobile which can capable of running a web browser and laptop with at least 8GB Ram or Nvidia GPU because I have used CUDA toolkit which supports only in Nvidia GPU or CPU. The software consists of Python, Django (Backend), Html, CSS, and JavaScript.

Required Libraries:

- Django
- Matplotlib
- NumPy
- Pandas
- PyTorch
- Spacy

Front End Technologies:

- HTML
- CSS
- JavaScript

Back End Technologies:

- Python
- Django Framework

PROCEDURE:

Steps that I followed to create a model:

1. Collected JSON data from Kaggle
2. Converted JSON data into a data frame for more convenience.
3. Cleaning the data.
4. Applied pre-processing techniques
5. Applied different Deep Learning algorithms and check the accuracy and other metrics by changing the parameters.
6. Picked the best model and deployed it in localhost using Django.
7. Created a Web Ui with Html, CSS, JavaScript.
8. Integrated the Web Ui with the backend.

INTEGRATION:

Django framework is used here to build a communication between python modules and with my website. The HTML page sends the news text to Django by using the post method. Then data goes to the deployed model the model gets processed and gives output

```
System check identified no issues (0 silenced).
January 02, 2022 - 18:19:56
Django version 4.0, using settings 'NewsClassifier.settings'
Starting development server at http://127.0.0.1:80/
Quit the server with CTRL-BREAK.
Not Found: //
[02/Jan/2022 18:20:12] "GET // HTTP/1.1" 404 2454
Not Found: /favicon.ico
[02/Jan/2022 18:20:13] "GET /favicon.ico HTTP/1.1" 404 2484
[02/Jan/2022 18:20:18] "GET / HTTP/1.1" 200 5085
[02/Jan/2022 18:20:19] "GET /static/abcd.png HTTP/1.1" 200 47386
[02/Jan/2022 18:20:19] "GET /static/bk1.jpg HTTP/1.1" 200 603627
[02/Jan/2022 18:34:27] "POST /predict HTTP/1.1" 200 39
```

OUTPUT SCREENS:

Sample 1

Sample 2


newsclassifier
- powered by NLP
Once again, Rajamouli sir Showing that why is he called LEGEND, I am from Mumbai and belive me, in Mumbai and All over East, West And North India, the craze of this movie is not less then South or even more.
Rajamouli Sir, You are really masterpiece
**Note: Transformers with CNN model is applying here..
Get Result


newsclassifier
- powered by NLP
ENTERTAINMENT
OK
Get Result

CONCLUSION:

The Hybrid model comprises two integral components: transformers and CNN. Initially, the transformers are employed to process and extract contextual information from the input text. Transformers, known for their proficiency in capturing long-range dependencies and contextual relationships within sequential data, play a pivotal role in understanding the nuanced semantics of news events. The output from the transformers, enriched with contextual representations, serves as the input for the subsequent CNN layer.

The CNN component of the Hybrid model is designed to further refine the extracted features. Convolutional layers are adept at capturing local patterns and hierarchical representations within the input data. By integrating the transformer output into the CNN architecture, the model gains the capability to discern both global contextual information and local intricacies, thereby enhancing the overall accuracy and effectiveness in identifying key features of news events. This innovative Hybrid model, combining the strengths of transformers and CNN, stands as a robust solution to the challenges posed by text classification and feature extraction within the dynamic realm of news event analysis.

The accuracy with Transformers+ CNN is 76.2%, and with CNN is 67.7%

REFERENCES:

1. NLP techniques for text classification:
   https://www.analyticsvidhya.com/blog/2021/12/text-classification-of-news-articles/

2. Applications of Deep Learning in News Text Classification:
   https://www.hindawi.com/journals/sp/2021/6095354/

3. Understanding the pre-processing techniques in details:
   https://www.youtube.com/watch?v=FLZvOKSCkxY&list=PLQVvvaa0QuDf2JswnfiGkli

4. Research paper on news classification:
   https://www.researchgate.net/publication/303501815_News_Classification_using_Neural

5. Word Embedding based News Classification by using CNN:
   https://ieeexplore.ieee.org/document/9537091

6. News Classification using Transformers:
   https://www.google.com/url?sa=t&source=web&rct=j&url=https://arxiv.org/pdf/2109.097

7. Labelling Text Data for News Article Classification and NLP
   https://omdena.com/blog/labeling-text-data-for-news-article-classification-andnlp/?doing_wp_cron=1656125526.5515940189361572265625