

Handy NewsRoom For Editors

DISSERTATION

Submitted in partial fulfillment of the requirements of
M. Tech Data Science Engineering Degree Programme

By

Aparana Bhatt
2018AD04510

Under the supervision of

Ashish Kanchan

Sr Engineering
Manager

Dissertation Work Carried At

New Delhi

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE PILANI (RAJASTHAN)

May, 2021

Handy NewsRoom For Editors

DISSERTATION

Submitted in partial fulfillment of the requirements of
M. Tech Data Science Engineering Degree Programme

By

Aparana Bhatt
2018AD04510

Under the supervision of

Ashish Kanchan

Sr Engineering
Manager

Dissertation Work Carried At

New Delhi

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE PILANI (RAJASTHAN)

May, 2021

ACKNOWLEDGEMENT

I would like to thank my mentor Ashish kanchan , for the immense support and suggestions he has provided me during the course and the project completion . Also his understanding and being helpful in my engagement to college work has been very fruitful for me and in turn played a key role in project completion .

Also I would like to extend my sincere gratitude to my examiner Ramesh Ramani to be of extreme help in the discussions . His suggestions provided me with a boost in the approach of completion of the project .

Last but not the least , my family members who have supported me immensely in every way they could and managed everything where I was needed but weren't present . Without this support project completion was not possible .

Aparana Bhatt

**BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE ,
PILANI**

CERTIFICATE

This is to certify that the Dissertation entitled 'Handy NewsRoom For Editors' and submitted by Aparana Bhatt ID. No. 2018AD04510 in partial fulfillment of the requirements of DSE CL ZG628T Dissertation, embodies the work done by him under my supervision.



Place: New Delhi

Signature of the Supervisor

Date : 6 August , 2021

Ashish kanchan

Sr Engineering Manager

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI II SEMESTER 2020-21
DSE CL ZG628T DISSERTATION
Dissertation Outline

BITS ID: 2018AD04510

Name of Student: Aparana Bhatt

Name of Supervisor : Ashish Kanchan

Designation of Supervisor : Sr Engineering Manager

Qualification and Experience: MCA 12+ Years of Industry Experience

Email Id of Supervisor : kanchan.ashish@gmail.com

Topic of Dissertation : Handy News-Room For Editors

Ramesh Ramani

Name of First Examiner: _____ **Designation of**

First Examiner: _____ **Qualification and Experience:**

rramani@wilp.bits-pilani.ac.in

E-mail ID of First Examiner: _____

Name of Second Examiner: _____ **Designation of**

Second Examiner: _____ **Qualification and Experience:**

_____ **E-mail ID of Second Examiner:**



(Signature of Student)



(Signature of Supervisor)

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI

**Work Integrated Learning Programmes Division
II SEMESTER 2020-21**

DSE CL ZG628T DISSERTATION

(EC-2 Mid-Semester Progress Evaluation Sheet)

Scheduled Month May:

NAME OF THE STUDENT: Aparana Bhatt

ID No. : 2018AD04510

Email Address : 2018ad04510@wilp.bits-pilani.ac.in

NAME OF SUPERVISOR : Ashish kanchan

PROJECT TITLE : Handy News-Room For Editors

EVALUATION DETAILS

EC No.	Component	Weightage	Comments (Technical Quality, Originality, Approach, Progress, Business value)	Marks Awarded
1	Dissertation Outline	10%		
2.	Mid-Sem Progress Seminar Viva Work Progress	10% 5% 15%		

	Supervisor	Additional Examiner
Name	Ashish Kanchan	
Qualification	MCA	
Designation & Address	Sr Engineering Manager, Times Internet Ltd	
Email Address	kanchan.ashish@gmail.com	
Signature		
Date	6/August/2021	

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI
Work Integrated Learning Programmes Division
II SEMESTER 2020-21

DSE CL ZG628T DISSERTATION

(Final Evaluation Sheet)

NAME OF THE STUDENT: APARANA BHATT

ID NO. : 2018AD04510

Email Address : 2018ad04510@wilp.bits-pilani.ac.in

NAME OF THE SUPERVISOR: ASHISH KANCHAN


PROJECT TITLE : Handy NewsRoom For Editors

(Please put a tick (✓) mark in the appropriate box)

S.No.	Criteria	Excellent	Good	Fair	Poor
1	Work Progress and Achievements	✓			
2	Technical/Professional Competence		✓		
3	Documentation and expression	✓			
4	Initiative and originality	✓			
5	Punctuality		✓		
6	Reliability		✓		
	Recommended Final Grade	✓			

EVALUATION DETAILS

EC No.	Component	Weightage	Marks Awarded
1	Dissertation Outline	10%	
2	Mid-Sem Progress		
	Seminar	10%	
	Viva	5%	
	Work	15%	
	Progress		
3	Final Seminar/Viva	20%	
4	Final Report	40%	
	Total out of	100%	

	Supervisor	Additional Examiner
Name	Ashish Kanchan	
Qualification	MCA	
Designation & Address	Sr Engineering Manager, Times Internet Ltd	
Email Address	kanchan.ashish@gmail.com	
Signature		
Date	6/August/2021	

Address: Aparana Bhatt D/O Virendra Kumar Bhatt

A 515-516 Eldeco Udyan 2 ,Ashray ,Near Delhi Public School RaeBareli Road

Lucknow

Pin Code: 226025

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI II
SEMESTER 2020-21
DSE CL ZG628T DISSERTATION

Supervisor's Evaluation Form

Supervisor's Rating of the Technical Quality of this Dissertation Outline

EXCELLENT/ GOOD/ FAIR/ POOR (Please specify): __GOOD__

Supervisor's suggestions and remarks about the outline (if applicable).

6/August/2021
Date _____



(Signature of Supervisor)

Name of the Supervisor: Ashish Kanchan

Email Id of Supervisor: kanchan.ashish@gmail.com

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI II
SEMESTER 2020-21
DSE CL ZG628T DISSERTATION

Dissertation Title : **Handy News-Room For Editors**

Name of Supervisor : Ashish Kanchan

Name of Student : Aparana Bhatt

ID No. of Student : 2018AD04510

ABSTRACT

Good news consists of effective summary meaningful and related keywords along with good and related images . News editors consistently work to provide effective and well written news . Due to rapid increase and digitisation of data , it has become inevitable to increase the content quality and attractiveness . There is an increased competition among media houses , as to which provides fastest news , most read news every day . For a news to be on the top in today's tech savvy world , it should have appropriate tags , effective keywords so that it's rank increases on the web and view count increases as well .

Also since the demands of good content are constantly increasing , it becomes necessary to improve the news editing process . Since most of the raw news articles do not contain news summaries and associated relevant keywords , they have to be manually edited . This process is exhaustive, less effective . If news summaries are automated and highlighted keywords are automatically generated , then associated images can also be searched very easily and the article can be made very attractive and complete . Also this reduces effort per article that an editor has to put .

For this project , the key pipeline is to provide effective news article summary and associated keywords for the given text and also provide associated image suggestions to the editor to create an effective content . The models will be trained on dummy data for articles of a specific category , once trained , they will be used to predict article summary and keywords. Then keywords will be used to fetch associated image suggestions by using available image search API's . Hence providing a workable and fast solution to create effective news content rapidly with reduced error probability .



(Signature of Student)



(Signature of Supervisor)

Tables

Table 1. Project Plan.....19

Contents

Chapters

1. Introduction & Background.....12

2. The Problem Statement.....13

3. Objectives of the Project.....14

4. Uniqueness of the Project.....15

5.Benefit to the Organization.....15

6. Scope of Work.....15

7.Solution Architecture.....16-17

8.Resources Needed for the Project.....18

9.Work Accomplished20-22

11.Key Challenges Faced.....23

12.References.....24

1 Introduction & Background

Good news consists of effective summary meaningful and related keywords along with good and related images . News editors consistently work to provide effective and well written news . Due to rapid increase and digitisation of data , it has become inevitable to increase the content quality and attractiveness . There is an increased competition among media houses , as to which provides fastest news , most read news every day . For a news to be on the top in today's tech savvy world , it should have appropriate tags , effective keywords so that it's rank increases on the web and view count increases as well . Over the years with advancement in intelligent technology and machine learning approaches , there have been various tasks to understand natural language processing , and its use cases .

For example generating effective text summary by using extensively trained deep learning models . Various approaches are available to derive text summarisation from a given trained model , using custom word embeddings like Glove, BERT etc . Many highly efficient models are available . Some of the models are T5,BERT . These models can be evaluated easily on test data and compared based on the industry use case and can be fine tuned , to meet our specific requirements .

Similarly there are various natural language processing based options available to extract meaningful keywords from a given text . Some of these options are TextRank,Dependency Parsing based approaches, RAKE etc .

This project aims at exploring multiple available models , comparing and finetune the models as per the requirement so as to produce effective summary and related valid keywords by using transfer learning approach primarily to achieve the outcome and provide an out of the box unified editor newsroom solution for effective and fast content creation . This solution also aims to provide related images by using generated keywords and associating with the articles so that editors get full fledged content at one place .

2 The Problem Statement

There is a newsroom content management system for news editors where they can access raw news articles provided by external agencies like PTI, ANI, REUTERS . Since media information needs to be precise and has time constraints related aspects . It becomes really important for editors to be able to edit and author impressive articles in minimum time . A good article consists of a good summary , relevant images and important keywords which make it visible and decide its rank on online search platforms like google . The raw stories provided by news providers many times do not contain proper summary,essential keywords and associated images which have to be manually inserted by the editors . For searching related images multiple keywords have to be searched for images . This all is a manual and time taking task of the current system . As a result, the latest news publishing becomes a slow process . The problem statement is to provide a comprehensive machine learning empowered solution to editors which alleviates the problem of manually writing article summary , writing keywords and then typing in multiple keywords to search the images from the internal image store and then embed it to the articles to make it publish worthy . The solution to this problem should save time as well as increase per editor article authoring capacity . It also aims to increase business value resulting from the enhanced content creation capacity.

3 Objectives of the Project

There are 3 main objectives that are targeted in the project

- a) **Time saving**
- b) **Empower Newsroom**
- c) **Relevant Image Availability**

Time Saving :

Articles without synopsis or summary need to be modified and editors have to write the synopsis manually , which is a time taking process . With the availability of synopsis on the go . This time will be saved and hence increases the throughput of the editors .

Empower NewsRoom :

The objective is to enrich the newsroom by automated content modification , prefill summary of the articles not having it , tagging content with appropriate keywords and also ensuring complete content provision for editors .

Relevant Image Availability :

Images add interactiveness to the news and using suggested keywords and then searching images across the available image corpus within the organisation's data ensures that images are appropriately tagged with the content . Also to reduce the effort of editors to search images manually , the aim is to search the images automatically and suggest them in the newsroom .

4 Uniqueness of the Project

Proposed project provides an inline comprehensive solution and ease of access of relevant data , which was otherwise invisible to editors , since it was dependent on manual search queries . The solution is a combination of different machine learning approaches solving two different sets of problems at once . It makes it very useful and can be projected as an in-house editorial tool for effective user experience .

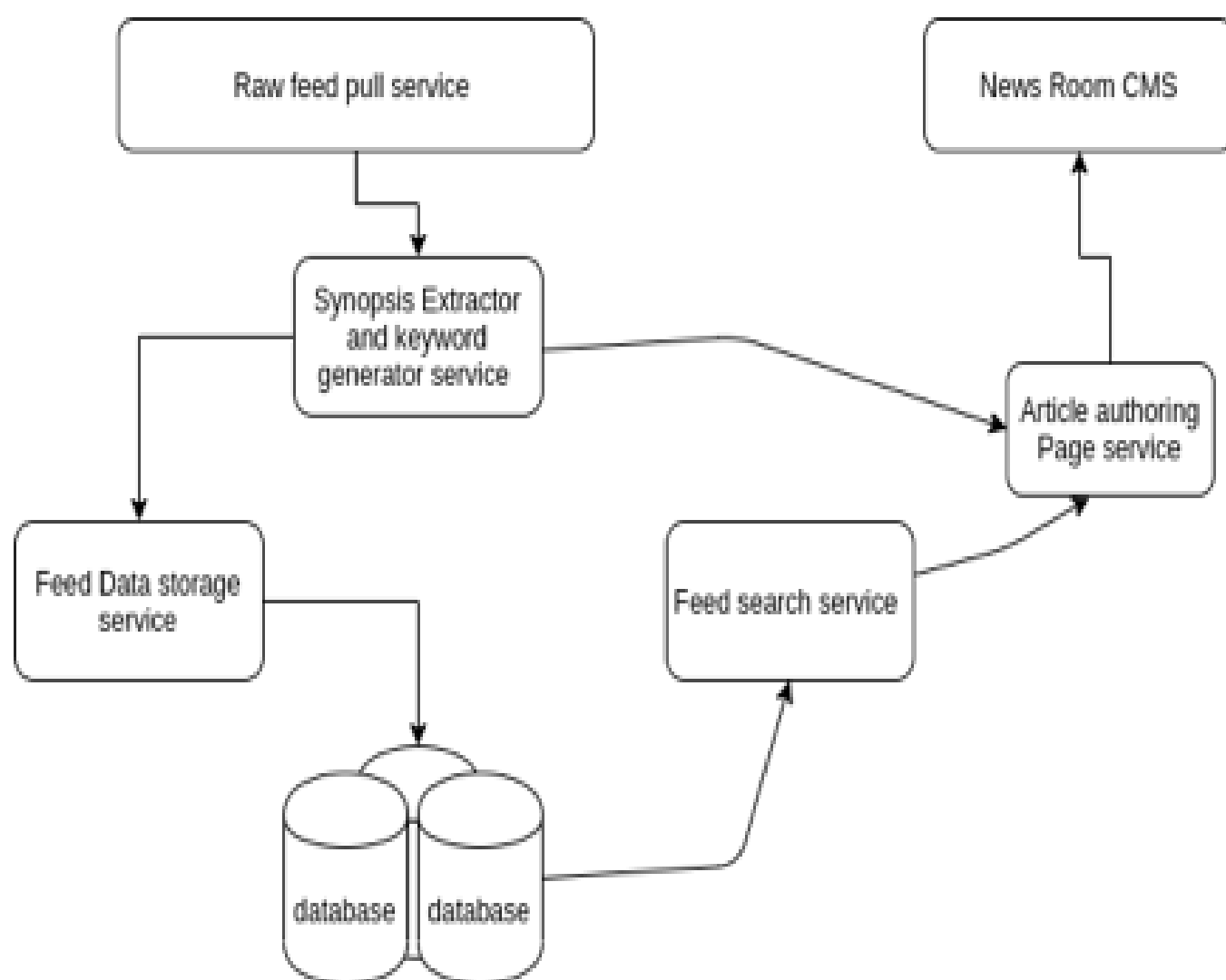
5 Benefit to the Organization

It enhances overall editorial capacity and also enables more efficient and relevant data searches . Relevant increase in usage of data for publishing is expected because more of it is available readily in fewer clicks and manual efforts .

6 Scope of Work

This project aims at exploring all the different ways to generate meaningful and relevant keywords as well as meaningful summary of articles by using machine learning algorithms and then creating apis to support its integration at the content management system level .

7 Solution Architecture



The architecture of the project is divided into 2 parts. Feed Storage Service and Feed Search Service

7.1 Feed Storage Service :

- As shown in the architecture diagram above, the feed storage service contains processed news articles at the time of ingestion into the newsroom storage system .
- Feed storage service , depends on **synopsis extractor and keyword extractor service** . Synopsis extractor and keyword extractor services are machine learning model based services that extract appropriate summary and relevant keywords from the input text and then pass to Feed Storage Service .
- For storing the text data , solr will be used . Solr provides various options like keyword search and also provides fast search of indexed text data .

7.2 The Feed Search Service:

- This service will be used to search the text data on the basis of search keywords .
- The results will be rendered from the solr corpus .
- Each text result will be prepopulated with autogenerated summary and keywords .
- There will be an option of related image search , when an article is selected for viewing .
- The complete article view will be available to the editor to publish the article .

8 Resources Needed for Project

Tools:

- Google Colab
- Apache solr
- Google Image search Api
- Pandas Library
- Scikit-Learn Library
- Matplotlib
- Transformers Library
- Visualization Tools
- Rogue Score Generator

Languages:

- Python

9 Project Plan & Deliverables

#	Task	Expected date of completion	Names of Deliverable
1	Data collection , preprocessing and model selection and comparison reports generation	May 5 , 2021	POC and ML modelling
2	Demo data indexing, and backend apis for searching texts , images , api for keyword generation , Summarization	May 15 ,2021	News view deliverables
3	Mid semester report generation	July 15-20,2021	Mid sem report
4	Inline on demand support api for Summarization and Keyword extraction api and basic article authoring page for editor	July 22-July 31,2021	UI integration
5	Complete demoable UI with complete working summarisation, keyword extraction and image search functionality	August 7, 2021	UI integration
6	Final project report and presentation	August 10th	Final report

Table 1. Project Plan

10 Work Accomplished So Far

10.1 Analysis and understanding of the news data for summarisation

Understanding various approaches and machine learning approaches available for text summarisation .

Exploring supervised and unsupervised algorithms and pretrained models available for the text summarisation .

Reading of published paper on transfer learning of T5 model , analysing the summaries produced by extractive summarisation and abstractive summarisation approach .

Reading of various blogs available for carrying out text summarisation for various texts .

Comparing effectiveness of generated summaries from reference summaries by ROGUE score .

10.2 Comparison of Various available text summarisation models

The main task was to analyse , preprocess and access the performance of various available models and compare their performance for a given set of custom input texts .

Currently various extractive text summarisation and abstractive text summarisation models are being explored . Main focus is to compare the existing pretrained models involving LSTM with attention models ,and BERT models .

For transfer learning on custom dataset , There will be comparison on model performances based on BERT embedding and GLOVE embedding .

For demo purposes of image search , google image search api has been explored
ROGUE score basis report generation flow has been made , which produces comparative report on generated summaries .

Once the model for both keyword extraction and text summarisation is finalised , then the data ingestion and search flow will be integrated .

Currently explored data models are T5,GPT2,BERT for abstractive summarisation and TextRank,KLSum,LuHn,LSA summarisers for extractive summary have been explored .

10.3 Analysis and understanding of the news data for keyword generation

Just like various approaches for text summarisation , there are multiple approaches for keyword extraction . The approaches can be classified as linguistic , statistical , machine learning based . Statistical approaches do not need training . They are based on word frequency , n-grams etc .

Linguistic approaches use complex analysis of language to devise keywords from a given text .

Machine learning approaches are classified as supervised , unsupervised , semi supervised and reinforcement learning .

As machine learning methods, which also use artificial neural networks, deep learning have achieved high performance in many areas. Deep learning utilises multiple layers of neural network and hidden layers . Researchers have also tried with the recurrent neural network (RNN) to extract topic keywords from different scales of texts and documents . Long **STM** (LSTM), **a specific sort of RNN**, achieved **an honest** performance in topic keyword extraction from many domain contents **for various** purposes. It differs from traditional RNN, and is **compatible** to classify, process, and predict text problems. However , the results are dependent on the type of datasets being used .

10.4 Custom training finalised ML approach for keyword generation

I have created BERT embeddings from the given sample text , to get a document level representation . These word embeddings are then extracted for N-gram words/phrases . These embeddings are used to train BERT models for keyphrase generation .

Then in the next step cosine similarity is used to identify most similar keywords to the document . This acts as an extra filter for the keywords suggested from the model output .

10.5 Related Image Search Integration

This step is not associated with Model training but associated with the outcome of generated keywords at the last step .

I have tried to extract meaningful keywords from the text and then searching associated images from the available corpus .

For the project purpose I will be using google image search from a custom browser , but in the organisation we will be using the available corpus of images which is actively used .

This step acts as a completion in the content shown to the editor for editing and publishing .

10.6 Data Model

Solr Data Model

Field	Value
Title	Title of document
Subject	ML generated synopsis
Story	Actual Story content
ContentType	If its image data or text data
KeyWords	ML generated keywords

11 Key Challenges Faced

- For text summarisation available models are so many , so to decide on 2-3 models to compare and finalise was an important task .
- Natural Language Processing training tasks require extensive hardware requirements , so to ease the constraint selective data for a particular category was chosen .
- Feature Engineering: Another challenge was to decide the features which will be fed to the model input for higher accuracy of prediction.
- Keyword generation models rely heavily on dataset , so benchmarking the same is a challenge .
- When searching related images , associated legal issues with the image credits and linking with a provider is an issue which requires manual validation as well .

12 References

<https://machinelearningmastery.com/gentle-introduction-text-summarization/>

<https://towardsdatascience.com/text-summarization-from-scratch-using-encoder-decoder-network-with-attention-in-keras-5fa80d12710e>

https://humboldt-wi.github.io/blog/research/information_systems_1920/nlp_text_summarization_techniques/

<http://ceur-ws.org/Vol-2718/paper28.pdf>

<https://pypi.org/project/bert-embedding/>

<https://www.machinelearningplus.com/nlp/text-summarization-approaches-nlp-example/>

<https://www.analyticsvidhya.com/blog/2019/06/comprehensive-guide-text-summarization-using-deep-learning-python/>

<https://www.kaggle.com/sandeepbhogaraju/text-summarization-with-seq2seq-model>

<https://www.kaggle.com/nehaytamore/abstractive-summarizer-in-keras>

<http://jaitc.ki-it.or.kr/xml/24833/24833.pdf>

<https://github.com/MaartenGr/KeyBERT>

<https://medium.com/mlearning-ai/10-popular-keyword-extraction-algorithms-in-natural-language-processing-8975ada5750c>

<https://github.com/andybywire/nlp-text-analysis>

<https://ieeexplore.ieee.org/document/8819819>