Enriching Text Summarization: A Journey through Contextual Guidance and Multimodal Data

Thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science
in
Computational Linguistics by Research

by

Anshul Padhi
2018114013
anshul.padhi@research.iiit.ac.in

International Institute of Information Technology
Hyderabad - 500 032, INDIA
February 2024

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled "**Enriching Text Summarization: A Journey through Contextual Guidance and Multimodal Data**" by **Anshul Padhi**, has been carried out under my supervision and is not submitted elsewhere for a degree.

Date	Adviser: Prof. Vasudeva Varma

To my parents and friends for their constant support.

Acknowledgments

First and foremost, I would like to extend my deepest gratitude to my advisor, Prof. Vasudeva Varma, for his invaluable guidance, mentorship, and unwavering support throughout the course of this research. His insights and dedication to academic excellence have been a source of inspiration. I'm equally indebted to my other advisors during this journey, Prof. Balaji Vasan Srinivasan, Prof. Kamal Karlapalem, and Prof. Manish Gupta for their invaluable feedback and for constantly challenging me to push the boundaries of my research.

I would also like to express my sincere appreciation to my research partners, Tanmay, Sayar, Risubh, and Nikhil. Collaborating with such a dedicated team has been both a pleasure and a privilege. My journey was further enriched by my peers at iREL. I am thankful for the camaraderie and the countless interesting discussions with Sumanth, Bhavyajeet, Himanshu, and Anubhav and many others.

To my close-knit college friend group, Daddycated, thank you for the cherished memories, unwinding sessions, and the continuous encouragement. Your friendship has been a beacon during the rigorous academic journey.

Lastly, but most importantly, I owe a debt of gratitude to my family. To my parents, who have always been my pillar of strength and reservoir of wisdom, and to my brother, who has been my confidant and cheerleader, thank you. Your belief in me, even during moments of doubt, was the driving force behind every page of this thesis.

This journey at IIIT-Hyderabad has been an unforgettable chapter of my life, and I am grateful to all who have been a part of it.

Abstract

The digital age presents an overwhelming deluge of multimodal data, underscoring the imperative need for effective text summarization techniques. Such techniques transform vast amounts of textual and visual data into concise, comprehensible, and insightful summaries, facilitating information retrieval, comprehension, and decision-making. This thesis pioneers innovative strategies to enhance text summarization by employing various forms of contextual guidance and multimodal data, contributing significantly to the evolution of the field and offering a cohesive narrative that links these diverse yet interconnected areas of study.

The journey begins with an exploration of "Popularity Forecasting" of sentences within news articles. This novel approach surpasses traditional salience-based extractive summarization by predicting the 'popularity' or 'eye-catching' potential of sentences. We create a popularity dataset which contains news articles from CNN/DM[47] with their sentence-popularity score mapping. We create this by comparing sentences with the search queries for the particular article. Then we adapt trained extractive summarizers to perform regression tasks and predict the popularity of a particular sentence within a news article. The result is a ranking of sentences based on their popularity scores

Next, the research advances into the realm of "Multimodal Summarization," which synergizes textual and visual elements to create a more holistic summary. By pairing concise textual summaries with the most salient images from news articles, this technique delivers a richer and more comprehensive understanding of the content. In this work we also show that we can improve the accuracy of summarization models by using images to aid the summarization process. To do this we utilize visuolinguistic transformers like CLIP[54], OSCAR[36] to help in the interaction of the two modalities and we adapt general summarization models so that we can incorporate both textual and visual information in the summarization model

Building on the foundation of extractive summarization, and using the core logic from the multimodal summarization work the study then introduces "Guided Summarization." This innovative method uses salience scores of sentences, obtained from an extractive summarizer, to guide an abstractive summarizer. This symbiotic relationship between the two forms of summarization results in more contextually relevant and focused abstract summaries.

The research further pushes the boundaries of personalization with "Persona-based Summarization," applied to SEBI legal case files. This technique generates tailored summaries

based on the specific information needs of different personas such as investors, defense lawyers, and judges. It underscores the potential of personalization in text summarization, making the information more accessible and relevant to each user profile.

Finally, building on the insights gleaned from the exploration of multimodal summarization, the study culminates with the creation of an "Indic Multimodal Text-Image Pair Dataset." This unique resource is a rich assembly of text and image pairs of different Indian languages, serving as a critical foundation for the development and evaluation of visuolinguistic transformers, especially those focusing on data from the Indian subcontinent.

In summary, this thesis provides a comprehensive exploration of how contextual guidance and multimodal data can significantly enhance text summarization. The innovative techniques and resources proposed and developed in this research, connected through a cohesive narrative, promise to significantly advance the field of text summarization, paving the way for more engaging, comprehensive, and personalized summary generation.

Contents

List of Figures

List of Tables

Chapter *1*

Introduction

1.1 Overview

The primary aim of this thesis is to explore and enhance the methodologies of text summarization. It seeks to achieve this by innovatively integrating various forms of contextual guidance and multimodal data into the summarization process. By conducting a series of interconnected studies that address the appeal, comprehensibility, relevance, and personalization of summaries, the thesis aims to contribute significantly to the evolution of text summarization techniques. Furthermore, it intends to provide a unique resource for future research in multimodal summarization through the creation of an Indic multimodal text-image pair dataset. Ultimately, this work endeavors to advance the field of text summarization towards more engaging, comprehensive, and personalized summary generation.

This thesis is divided into the multiple works that I have done on summarization. Each chapter contains one research work that tries to enrich or adapt text summarization using various methods and resources.

1.2 Popularity Forecasting

The digital landscape has transformed the way we consume and disseminate information, particularly in the realm of online news. A critical aspect of this transformation is the ability to predict the popularity of individual sentences within these news documents, a task that forms the crux of our study. This task is not merely an academic exercise; it holds significant implications for various sectors, including journalism, marketing, and social media, where the popularity of information can influence public opinion, drive consumer behavior, and shape societal trends.

The current research landscape on this topic is still developing. While there have been strides in related areas such as text summarization and snippet generation, the task of sentence-specific popularity forecasting is relatively uncharted territory. This task presents unique challenges,

as it requires a nuanced understanding of natural language content and the ability to predict internet browsing behavior, both of which are complex and dynamic phenomena.

In the context of this thesis, this work contributes by introducing a new dimension to text summarization. By predicting the popularity of individual sentences, we can enrich text summarization models with an additional layer of contextual guidance. This can potentially lead to more effective and relevant summaries, as the models can prioritize information that is likely to be popular or impactful.

In this work, we aim to contribute to the research on this task by introducing a novel dataset, InfoPop, which contains popularity labels for over 1.7 million sentences from over 50,000 online news documents. Leveraging this dataset, we propose a novel transfer learning approach that uses sentence salience prediction as an auxiliary task and a BERT-based neural model. Our approach aims to enhance the prediction of sentence popularity by learning from salience prediction, thereby bridging the gap between what is important in a document (salience) and what is likely to be popular among readers.

1.3 Text-Image Multimodal Summarization

This work explores the realm of multimodal summarization, a burgeoning field that aims to condense information from multiple modalities, such as text and images, into a concise and interpretable form. The task at hand is the development of a multimodal summarization model, specifically designed for news articles, that leverages semantic reranking and cross-modal knowledge distillation.

The importance of this task is underscored by the rapidly growing amount of multimodal content available on the internet, particularly in news articles and blog posts. The ability to transform such articles into concise and accessible formats, akin to social media posts, presents several noteworthy advantages for information dissemination, including wider reach, better retention, and easier access to content.

The current state of research in this area has seen efforts to solve the problem of condensing such articles into summaries by using attention mechanisms to pool the modalities together. However, these efforts have not been entirely successful in forming an effective representation. Previous models, such as MSMO and UniMS[76], have utilized techniques like bidirectional LSTM[51] and BERT's encoder-decoder architecture, but they have not fully exploited the potential of multimodal data.

In this work, we aim to address these shortcomings by proposing a knowledge-distillation based approach that manages to separate textual content of higher quality through contrastive learning utilizing pretrained multimodal models. Our model, MMSumm, uses BERT[18] for generating contextual embeddings of the text and VGG-19[60] for the images, and it employs a scoring mechanism to determine the semantic similarity of the image to the text.

1.4 Salience Guided Summarization

The task of text summarization has seen significant advancements, yet there remains a considerable scope for improvement. One such area of exploration is the use of intrinsic text-based measures to enhance the performance of abstractive summarization. This forms the basis of my current research, where I am investigating how these measures, such as an extractive summarization or a set of keywords and key phrases extracted from the text, can be used to guide the abstractive summarization process.

The motivation behind this work stems from the understanding that while abstractive summarization has the potential to generate more coherent and concise summaries, it can often miss out on key information or deviate from the original text's meaning. By incorporating intrinsic text-based measures as guidance, the aim is to create a balance between abstraction and accuracy, thereby boosting the overall quality of the generated summaries.

Existing work in this area includes the GSUM[19] model, which incorporates different forms of guidance to enhance abstractive summarization. While the initial goal of my research was to create a model that could outperform GSUM, this has proven to be a challenging task. However, the exploration of different ways to incorporate guidance into the summarization pipeline has yielded valuable insights and has opened up new avenues for further research.

In the context of my thesis, "Enriching Text Summarization: A Journey through Contextual Guidance and Multimodal Data", this work contributes significantly. It aligns with the central theme of the thesis, which is to enrich text summarization through contextual guidance. The exploration of intrinsic text-based measures as guidance is a form of contextual guidance, and the insights gained from this research can inform and enhance the methodologies explored in the thesis.

1.5 Personalised Summarization

Legal case files often present a significant challenge due to their complexity and volume of information. Various stakeholders, such as investors, defense lawyers, and adjudicating officers, often find it difficult to swiftly extract relevant information. This challenge serves as the motivation behind the research paper titled "Aspect-based Summarization of Legal Case Files using Sentence Classification".

Previous attempts to address this problem have employed machine learning and deep learning methods for multi-class classification of sentences in legal documents. Techniques such as BERT[18] for sentence classification and hierarchical attention networks for document classification have been explored. Furthermore, methods for abstractive and extractive summarization, such as LexRank[21] for salience in text summarization and sequence-to-sequence RNNs for text

summarization, have been utilized. However, these methods often do not consider the specific needs of different stakeholders.

This research aims to fill this gap by developing a method for aspect-based summarization of legal case files. The goal is to generate summaries tailored to the specific needs of different stakeholders, thereby making the information in legal case files more accessible and useful.

The relevance of this research to my thesis, "Enriching Text Summarization: A Journey through Contextual Guidance and Multimodal Data", is profound. It provides a practical demonstration of how contextual guidance can be employed to enhance text summarization, a central theme of my thesis. The aspect-based summarization approach presented in this paper uses the context of the legal case and the specific needs of different stakeholders as a form of contextual guidance to direct the summarization process. This approach aligns seamlessly with the theme of my thesis and serves as a valuable reference for the methodologies explored in my work.

1.6 Indic Multimodal Work

The richness and diversity of Indian languages present a unique opportunity and challenge for the field of text summarization. As part of my ongoing research, I am currently working on the creation of an Indic multimodal dataset. This dataset, comprising 50,000 sentences in various Indian languages along with relevant photos extracted from Google, is a pioneering effort in the field.

My motivation for this work is to further the development of an Indic visuolinguistic transformer akin to OpenAI's CLIP[54]. This is a significant endeavor, as there is currently no authentic Indic multimodal dataset available. Existing datasets are largely artificial, often translated or transformed from English multimodal datasets, which do not fully capture the nuances and richness of Indian languages.

The incorporation of this work into my thesis, "Enriching Text Summarization: A Journey through Contextual Guidance and Multimodal Data", is a natural extension of the research. Currently, my work in multimodal summarization is limited to English news articles. By creating an Indic variant of CLIP, I aim to extend this work to Indic languages. This would involve using the Indic variant of CLIP to rate sentence similarity with article photos and training the rest of the summarization model on Indic text.

This endeavor contributes significantly to my thesis by expanding the scope of multimodal summarization to include Indic languages. It also aligns with the central theme of my thesis, which is to enrich text summarization through contextual guidance and multimodal data. By creating a new dataset and developing an Indic visuolinguistic transformer, I am adding a new dimension of context - that of language diversity - to the field of text summarization. This

work serves as a valuable reference for the methodologies explored in my thesis and paves the way for future research in this area.

1.7 Contributions of this Thesis

This section summarizes the key contributions of this thesis:

1. Proposing the novel task of sentence level popularity prediction for news articles, alongside a dataset and a
2. Proposing a novel architecture for image-text summarization of English news articles where both modalities guide each other in creating the best overall summary
3. An exploration in my attempts to incorporating text based guidance from within the text to boost summarization performance
4. Proposing a method for generating personalised summaries for different class of people based on their requirements in the legal domain
5. Creating and contributing a high quality multimodal dataset in Indic languages.

1.8 Thesis Workflow

This thesis is divided into seven chapters. The first chapter is the introduction of the thesis, the second chapter contains an overview on the the existing work done on tasks that are relevant to the work persented in this thesis. Chapters three to seven cover the different works done as part of this thesis. Chapter eight concludes the thesis by summarizing the contributions and discussing possible future works related to this thesis.

- Chapter 1 aims to provide a general overview of the thesis. It has introductory information on text summarization, along with an overview of the different types of summarization and ways to enhance text summarization
- Chapter 2 describes the existing literature that is relevant to the work presented in this thesis.
- Chapter 3 covers *SCATE*, which is one of the first works in its domain. It includes the overall work and its scope, along with the dataset extraction process and the technical architectural details of the model that we propose for this task.
- Chapter 4 covers *MMSumm*, which is a text-image multimodal summarization model. We discuss the motivation and the architecture of our model along with the results.

- Chapter 5 covers my exploration of different forms of intrinsic text based guidance and how to best incorporate them into the summarization model architecture to boost summarization performance over vanilla text summarization
- Chapter 6 covers *Finweb*, which is an approach for persona based text summarization for SEBI legal case files. We discuss the approach, and also discuss how this can be useful for other domains.
- Chapter 7 covers the creation of our Indic Multimodal dataset. This is a text-image dataset for Indic languages. We discuss the scope and motivation, along with the issues we faced along the way and how we handle the issues.
- Chapter 8 concludes this thesis with a summary of our contributions along with potential ideas that can be explored further.

Chapter *2*

Related Work

This chapter aims to give an overview on the major works done in the field of text summarization along with the works that are related to the specific tasks that I have worked on for this thesis

2.1 Text Summarization

Text summarization, a subfield of Natural Language Processing (NLP), aims to generate a concise version of a text while preserving its key information and overall meaning. This process can be broadly categorized into two types: extractive summarization and abstractive summarization.

2.1.1 Extractive Summarization

Extractive summarization, one of the earliest approaches to text summarization, selects key sentences or phrases from the source text to form the summary. The objective is to identify and extract the most informative segments of the text without altering the original wording.

One of the first works in extractive summarization was way back in the 1950s with Luhn's algorithm[1], proposed by Hans Peter Luhn. This algorithm identifies the most frequent non-stop words in a document and selects sentences with the highest frequency of these words.

In the subsequent years, the MEAD system[53] emerged, which uses cluster centroids for multi-document summarization. It computes the centroid of a document cluster as the mean word frequency vector of the documents and ranks sentences based on their similarity to the centroid.

The number of works in extractive summarization increased a lot when machine learning became popular. The TextRank algorithm[46], inspired by Google's PageRank[8], uses a graph-based ranking model to select sentences. One of the state-of-the-art models in extractive sum-

marization is BERTSumm[41]. This model extends the powerful BERT[18] language model to generate interval representations for sentences, thereby enhancing extractive summarization.

2.1.2 Abstractive Summarization

Abstractive summarization, is another type of summarization that generates new sentences to encapsulate the main ideas of the source text. This approach can yield more coherent and concise summaries but is generally more challenging due to the complexity of natural language generation.

Early works in abstractive summarization usually used rule-based methods, such as the use of semantic representations and sentence compression techniques. However, these methods were limited by the complexity and variability of natural language.

More sophisticated models were introduced with the rise in deep learning. One of the first such models was the sequence-to-sequence model with attention[59], which was used to generate abstractive summaries of news articles.

More recent works have used transformer-based models, such as BERT and GPT[9], for abstractive summarization. For instance, the PEGASUS model[73], developed by Google, uses a transformer model pre-trained on a large corpus of text and fine-tuned for the task of abstractive summarization.

One of the state-of-the-art models in abstractive summarization is BART[34] (Bidirectional and Auto-Regressive Transformers), developed by Facebook. BART is trained to reconstruct the original text by randomly masking out sentences from the text and then generating the masked sentences.

2.2 Popularity forecasting of news article sentences

The task of predicting the popularity of online news sentences is a relatively new but rapidly evolving field. Several notable studies have made significant contributions to this area.

Tatar et al. [64] was one of the first works that worked on the concept of ranking news articles based on their predicted popularity. Their work laid the groundwork for future research in this area by demonstrating the potential of popularity prediction as a tool for news article ranking.

Uddin et al. [65] expanded on this concept by exploring the use of content metadata for predicting the popularity of online news. Their approach highlighted the value of metadata as a rich source of information for popularity prediction.

Voronov et al. [67] took a different approach by focusing on the title of news articles. They employed a BN-LSTM network to analyze titles and forecast the popularity of the articles. This work underscored the importance of titles as a key factor in driving the popularity of online news.

Wang et al. [29] proposed a feature generalization framework for predicting social media popularity. While their work focused on social media, the principles of feature generalization they introduced are relevant to the broader field of online news popularity prediction.

Lastly, Wu et al. [69] introduced the concept of multi-scale temporal decomposition for predicting social media popularity. Their work emphasized the importance of considering temporal dynamics in popularity prediction.

These studies collectively represent the current state of the art in online news popularity prediction. They provide valuable insights and methodologies that inform and inspire our work on sentence-specific information popularity prediction.

2.3 Multimodal Summarization

Multimodal summarization, particularly the integration of text and image data, has been a burgeoning field of research in recent years. This section provides an overview of some of the key contributions in this area.

Jiang et al.[27] proposed a contrastive learning strategy that refines cross-modal similarity progressively. This strategy aims to optimize the mutual information between an image/text anchor and its negative counterparts more accurately. Their work provides valuable insights into the potential of contrastive learning in multimodal summarization.

Lu et al. introduced MTCA[45], a multimodal summarization model based on two-stream cross attention. The model comprises a pre-trained feature extractor, a text encoder, an image encoder, a two-stream cross attention fusion module, and a summary decoder. This comprehensive approach to multimodal summarization represents a significant advancement in the field.

Zhang et al.[74] proposed a hierarchical cross-modality semantic correlation learning model (HCSCL) to learn the intra- and inter-modal correlation existing in multimodal data. Their model outperforms the baseline methods in automatic summarization metrics and fine-grained diversity tests, demonstrating the potential of hierarchical learning models in multimodal summarization.

Lastly, Zhang et al. proposed UniMS, a unified framework for multimodal summarization grounded on BART[34]. UniMS[76] integrates extractive and abstractive objectives and includes a visual guided decoder to better integrate textual and visual modalities in guiding abstractive text generation. This work underscores the potential of unified frameworks in enhancing multimodal summarization.

These works collectively highlight the ongoing research and development in the field of text-image multimodal summarization. They underscore the potential of various approaches, from recurrent neural networks and attention mechanisms to contrastive learning strategies and unified frameworks, in advancing the field.

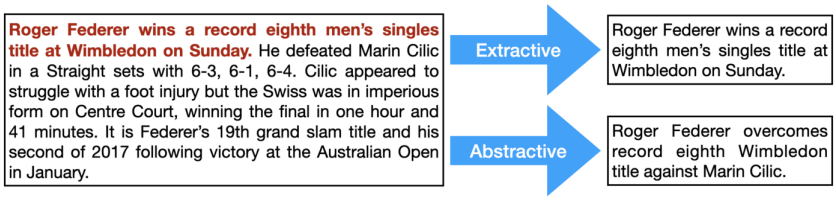

Figure 2.1 Examples of extractive and abstractive summarization.

2.3.1 Multimodal summarization with multimodal output

The domain of abstractive summarization encompasses the task of multimodal summarization with multimodal output (MSMO)[77]. This task involves leveraging image content within predominantly textual documents to enhance abstractive summarization. This concept was initially put forth by Zhu et al. [77], who developed a dataset comprising CNN and Daily Mail news articles and their associated images. Utilizing an attention-based sequence-to-sequence model constructed with bidirectional Long Short-Term Memory (LSTM) networks[23], they demonstrated remarkable performance for that period, as evaluated by human assessors. Their research also revealed that presenting the most pertinent image alongside a text-only summary significantly improved user satisfaction. In a subsequent study, Zhu et al. [28] further refined their model by incorporating a multimodal ranking method to rank images. Currently, the highest metrics on this task have been achieved by Zhang et al. [76], who employed BART[34] in conjunction with CLIP[54] as a knowledge distillation module.

2.4 Persona based summarization of SEBI legal case files

The field of aspect-based summarization has seen significant contributions over the years. This chapter provides an overview of the key works that have shaped this area of research.

Hu and Liu (2004)[25] pioneered the application of aspect-based summarization in the context of unstructured product reviews. They employed association mining to identify frequent word itemsets, which were then used to discern aspects and their associated sentiments. This work laid the foundation for subsequent research in the field.

Building on this, Angelidis and Lapata (2018)[4] introduced a convolutional neural network-based model for aspect-based summarization. Their approach involved a two-step process where aspects were first identified, followed by the generation of a summary for each aspect. This work demonstrated the potential of deep learning techniques in aspect-based summarization.

Kreimeyer et al. (2017)[32] applied aspect-based summarization to clinical study reports. Their approach combined natural language processing and machine learning techniques to identify aspects and generate summaries. This work underscored the potential of aspect-based summarization in the medical field.

Lastly, Lerman et al. (2009)[33] focused on aspect-based summarization of business news. They used a combination of natural language processing and machine learning techniques to identify aspects and generate summaries. This work demonstrated the applicability of aspect-based summarization in the business domain.

2.5 Guided text summarization

The field of guided summarization has seen significant contributions over the years. This chapter provides an overview of the key works that have shaped this area of research.

One of the early works in this area is "Topic-Focused Multi-Document Summarization Using an Approximate Oracle Score" by Conroy and O'Leary (2001)[14]. They proposed a method for guided summarization using Hidden Markov Models (HMMs). The HMMs were used to extract salient sentences from the document, which were then used as guidance for the summarization process.

Another notable work is "Fast Abstractive Summarization with Reinforce-Selected Sentence Rewriting" by Chen and Bansal (2018)[12]. They proposed a method for guided abstractive summarization using extractive summaries. The process involved generating an extractive summary first, which was then used as guidance for the generation of the abstractive summary.

A significant contribution to the field of guided summarization is "Guided Abstractive Summarization with Explicit Information Selection Modeling" by Zhao, Wang, and Neubig (2021). They introduced a new framework, GSum[19], which uses a guidance signal to control the structure of the generated summaries. The guidance signal can be any form of structured input, such as a set of extracted keywords, a list of entities, or an extractive summary. The authors demonstrate its effectiveness through experiments, achieving state-of-the-art performance on four popular summarization datasets when using highlighted sentences as guidance. The GSum model not only generates more faithful summaries but also provides a degree of controllability, demonstrating

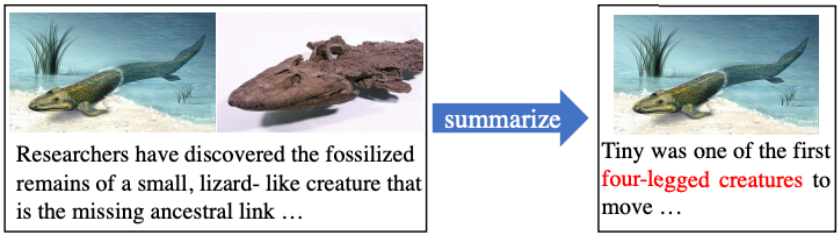

Figure 2.2 Example of multimodal summarization with multimodal output.

2.6 Multimodal Datasets

One of the early works in this area is the creation of the MSCOCO dataset[38] by Microsoft. This dataset, which includes captions for each image, has been widely used for image recognition tasks and has become a valuable resource for multimodal research.

Similarly, ImageNet[17], primarily an image recognition dataset, has been used in conjunction with other text datasets to train multimodal models.

Google's Conceptual Captions[58] is another large-scale dataset that contains images paired with captions, which are more descriptive and longer than those in MSCOCO.

In terms of visuolinguistic transformers, OpenAI's CLIP (Contrastive Language-Image Pre-training)[54] represents a significant advancement. CLIP jointly trains to understand images and text from a large dataset of internet text paired with images.

ViLBERT (Vision-and-Language BERT)[44] and VisualBERT[35] are other notable models that extend BERT to handle both images and text. They are trained on the Conceptual Captions[58] and MSCOCO datasets, respectively.

In the context of Indian languages, the creation of the "Hindi Visual Genome" dataset[49] is a significant milestone. This dataset, which is the first of its kind for English-Hindi multimodal machine translation, consists of short English segments (captions) from the Visual Genome, along with associated images. These segments have been automatically translated into Hindi, with manual post-editing that took the associated images into account.

In our attempt to get an Indic multimodal dataset, we make use of the samanantar dataset from AI4Bharat [55], which is text-only. Samanantar dataset consists of 49.7 Million pairs of sentences between English and 11 Indic languages spread across 2 language families - Indo-aryan and Dravidian. These languages are - Hindi, Bengali, Tamil, Telugu, Odiya, Kannada, Assamese, Marathi, Punjabi, Gujarati and Malayali.

Chapter *3*

Popularity Forecasting

In our study, we focused on predicting how popular individual sentences in online news articles would be based on their content. We created a dataset named InfoPop, which includes popularity scores for over 1.7 million sentences from more than 50,000 online news articles.

In our approach we used a technique called transfer learning, applying it to sentence salience prediction as a side task. We built a neural network model using BERT, a well-known language processing framework. This model has performed well, with nDCG scores over 0.8 in predicting the popularity of sentences.

Our results show that applying transfer learning from sentence salience prediction improves the accuracy of predicting sentence popularity.

3.1 Introduction

The digital revolution has reshaped how news is consumed, transitioning from traditional print and television to diverse online platforms. This shift has led to a surge in content, with news providers competing to engage readers in this information-saturated environment. In such a scenario, discerning what resonates with readers is essential. A key aspect of this is understanding the impact of individual sentences within a news article, as they significantly contribute to the overall narrative. Analyzing sentence popularity can offer insights into reader preferences and optimize content and ad placement. However, predicting sentence popularity is a complex task, prompting our research to develop a model to forecast the popularity of sentences in online news articles.

The primary objective of this work is to address the task of predicting the popularity of individual sentences in online news articles. We aim to develop a model that can accurately predict sentence popularity, providing valuable insights for news providers navigating the digital age. This chapter delves into our research, offering an overview of our approach and its potential implications. As the second author, my contributions were in the development and implementation of the model, as well as in creating the InfoPop dataset, which serves as the foundation

of our research. This chapter aims to provide an account of our research, highlighting our approach and its potential impact on news content analysis.

The major contributions of our paper are as follows:

- We introduce the task of proactively forecasting relative information popularities of sentences within online news documents solely based on their natural language content, without the use of any external features.
- We present InfoPop, the first labeled dataset containing over 50,000 online news documents from 26 news websites with over 1.7M sentences, each mapped to supervised popularity scores.
- Through a novel STILTs-based transfer learning approach, we build high-performance neural models reaching nDCG scores over 0.8 for sentence-specific popularity forecasting.
- These contributions not only advance the field of information popularity prediction but also provide valuable resources for future research in this area.

Sentence	**Popularity Score**
Local carriers and drivers will be able to book ...	0.1099
Uber Freight plans to expand to more Euro...	0.2046
The EU and U.S. freight markets have problem...	0.0092
They’ re both huge —the EU truckload marke...	0.0067
“The European trucking market is experienc...	0.0110
Inefficiency of this scale results in shippers ...	0.0132
Uber Freight has been scaling up its business ...	0.1356
The company has offices in San Francisco and ...	0.0158
In August, Uber announced that it would ...	0.2703
It’ s also made some key hires, one of which ...	0.0076

Table 3.1 Example of sentences extracted from a news article and their corresponding popularity scores.

3.2 Background

This work is a combination of many problems that are related to the topic at hand

Popularity Prediction at the Document Level: A significant portion of the existing research in popularity prediction has been centered around treating an entire piece of content as a single unit for which the future popularity is predicted[67, 5, 29]. This prediction often relies on two main types of popularity indicators. The first is based on internet browsing habits, where the popularity of online news articles is gauged by the number of page load requests they receive over a certain period.[3] The second type of popularity prediction is focused on social media engagement[63, 61]. This approach has been applied to a variety of content types, including multimedia posts, images, movies[**?**, 2], and petitions. The popularity in this context is usually measured using indicators of user behavior such as the number of comments or shares[65]. Some researchers have also explored the prediction of social media popularity over time, using time-aware and time-series prediction models[69].

Automatic Text Summarization: Text summarization aims to identify and highlight the central idea of one or more documents. There are two main approaches to this: abstractive and extractive. Abstractive summarization generates a concise, coherent summary that encapsulates the central idea of the text. Extractive summarization, on the other hand, selects and arranges the most salient and diverse parts of the text to form a summary. Unlike these methods, our task involves sequential regression, where we forecast numerical scores for each sentence.

Snippet Generation: Snippet generation[50] involves creating a brief excerpt from a document that allows a user to understand the relevance of the document to their query without having to access the entire document. While the problem formulation for snippet generation is somewhat related to sentence popularity forecasting, our task is focused on query-insensitive scoring of sentences.[6] In contrast, tasks like snippet generation and document retrieval are query-sensitive and require a specific query to generate an appropriate snippet.

These works have provided valuable insights and methodologies that have informed our approach to forecasting sentence-specific information popularity within online news documents.

3.3 Dataset

Our research relies on the InfoPop Dataset, comprising 51,770 news documents containing 1,711,890 annotated sentences. The dataset's characteristics include variable document lengths, averaging 33.07 sentences per article, with sentences containing an average of 18.23 word tokens. It features contributions from 26 reputable news websites, averaging 1991.15 articles per site. We observed a weak positive correlation (0.168) between popularity scores and sentence lengths.

We split the dataset into train, validation, and test sets in an 8:1:1 ratio. The dataset's creation involved scraping 82,540 news documents from 26 online news websites and addressing noise issues in the text, implementing two dependency parsing-based heuristics.

The first heuristic eliminated sentences with non-tree dependency graphs. The second heuristic removed sentences with xcomp branches leading to a single participle, linked to text rep-

etition. These computationally intensive cleaning steps were conducted once, followed by the removal of articles with fewer than three grammatical sentences.

To assign popularity scores, we considered document relevance to Bing Search queries, focusing on documents in the top 10 search results. Each sentence received a base score, normalized within each document to a [0, 1] range. This process resulted in a dataset offering a diverse collection of news documents with sentence-level popularity annotations based on real-world search engine queries.

3.4 Model

Given the dataset mentioned above we aim to create a model that can solve the following problem statement

Given a document D with n sentences $s_1, s_2, \ldots, s_n$, our goal is to predict a sequence of popularity scores $p_1, p_2, \ldots, p_n$ where each p_i represents the predicted popularity of sentence s_i.

The popularity score for each sentence is a real number in the range $[0, 1]$, and the sum of all popularity scores in a document equals 1, i.e., $\sum_{i=1}^{n} p_i = 1$. This is to ensure that the scores represent a distribution of popularity across the sentences in the document.

3.4.1 $Base_{Reg}$ Model: Comprehensive Overview

The $Base_{Reg}$ model is tailored for sentence scoring in online news documents and takes inspiration from the *SummaRunner* architecture[48].

3.4.1.1 Input Representation and CNN-based Vectorization

Sentences are initially represented using GloVe embeddings[51] and transformed into vectorized representations through a dual-layered CNN. This network employs convolution, batch normalization, and Leaky ReLU activation to capture localized patterns. A multi-kernel strategy with max-pooling is used to extract prominent features, yielding a comprehensive sentence embedding.

3.4.1.2 Contextualization with Bi-GRUs

The CNN-generated sentence vectors are refined using bidirectional Gated Recurrent Units (bi-GRUs)[13], ensuring that each sentence representation is influenced by its surrounding context.

3.4.1.3 Global Contextual Understanding

The bi-GRU-produced contextual embeddings are aggregated via max-pooling and processed through a fully connected layer, resulting in a document embedding vector that encapsulates the global context.

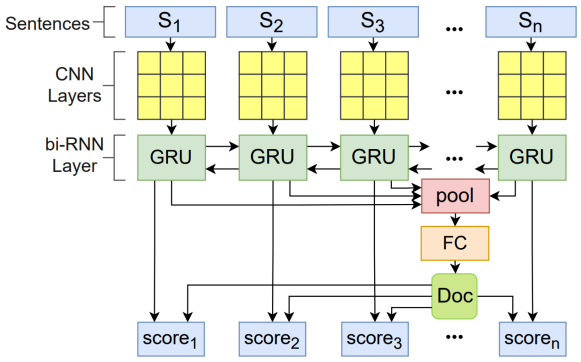

Figure 3.1 $Base_{Reg}$ Model

3.4.1.4 Sentence Scoring Mechanism

The culmination of the model's operations is the computation of scores for each sentence. These scores are derived based on various criteria, such as content richness, novelty, and other intrinsic sentence attributes. The scoring mechanism is a function of both the contextual sentence embedding and the global document embedding, ensuring a balanced consideration of local and global contexts.

3.4.2 $BERT_{Reg}$ Model: An In-depth Examination

The $BERT_{Reg}$ model, as delineated in the paper, represents an adaptation of Transformer-based architectures, specifically BERT[18], tailored for the task of sentence-specific popularity forecasting within online news documents.

3.4.2.1 Architectural Inspiration

The $BERT_{Reg}$ model harnesses the capabilities of BERT (Bidirectional Encoder Representations from Transformers) [18] to set new standards in multiple NLP benchmarks. This model modifies the *BertSumExt* [41] design, initially meant for sentence classification, to fit the sequential regression setting of the current task.

3.4.2.2 Sentence Embedding Generation

For contextual sentence embedding generation, the model marks sentence boundaries using [SEP] tokens and begins sentences with unique [REG] tokens. The [REG] tokens' contextual embeddings serve as the basis for sequence regression. The model then crafts embeddings for each input token, accounting for token, position, and segment. Notably, segment embeddings distinguish between odd and even sentences in BERT before being refined within the BERT framework. The sequence of BERT-informed sentence embeddings associated with [REG] tokens is merged with sinusoidal positional embeddings. Subsequently, they are passed through a two-layer Transformer model [66]. This results in contextual sentence embeddings, aware of the wider document context, emphasizing the model's comprehension of inter-sentence relationships.

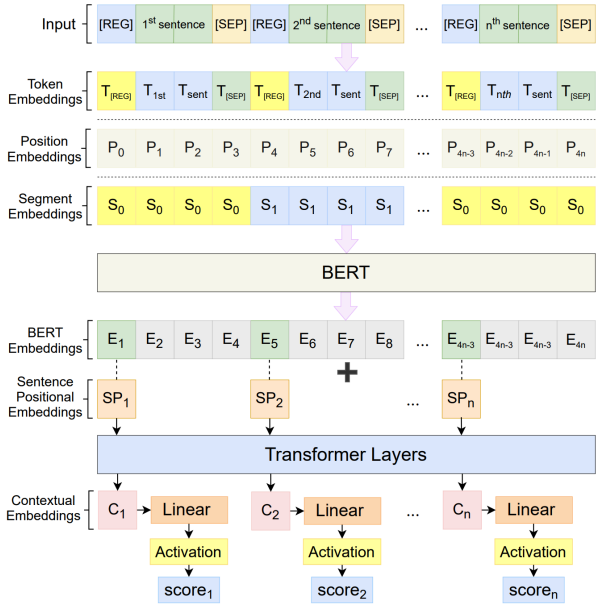

Figure 3.2 $BERT_{Reg}$ Model

3.5 Auxiliary Transfer Learning Subtasks

3.5.1 Conceptual Foundation

One of the primary innovations in our research approach is the incorporation of a STILTs-based Transfer Learning (TL) setup[20], focusing on the task of sentence salience prediction. We recognized that the domain of text summarization, which extensively studies sentence salience, offers a wealth of data from the news domain. Based on this observation, we posited that

leveraging STILTs-based transfer learning from an auxiliary task centered on salience prediction could significantly enhance the capability of models to forecast popularity[22].

3.5.2 Salience Prediction

In the realm of document summarization, a sentence's salience is gauged by its relevance to the core semantics of the document, determining its suitability for inclusion in the document' s summary. Given an article paired with its gold standard summary, We devised a method to compute the salience of each sentence based on its ROUGE[37] overlap with the summary. Traditional extractive summarization methods often assign binary summary-inclusion labels to sentences in a manner that maximizes the ROUGE overlap between the constructed oracle and the actual summary. However, for the purposes of my study, we aimed to capture the intrinsic salience of sentences, ensuring that lexically similar sentences received analogous labels. This approach led us to frame the auxiliary task as a sentence sequence regression problem, enabling the adaptation of the STILTs methodology and facilitating empirical cross-task evaluations.

3.5.3 Data Source for Salience Prediction

Our research primarily focuses on the online news domain. To this end, we utilized the well-known CNN-DailyMail news summarization dataset[47]. For each sentence, we computed three weakly supervised salience scores based on its ROUGE 1, ROUGE 2, and ROUGE L overlaps with the associated article' s summary. Similar to the methodology we employed for InfoPop labels, we normalized the salience labels across documents by dividing them by the cumulative score of individual sentences. This approach resulted in the formulation of three distinct auxiliary subtasks, labeled as $S1$, $S2$, and SL.

3.6 Experimental Details

3.6.1 Baselines

The paper employs a diverse set of baselines to benchmark the performance of the proposed models for sentence-specific popularity forecasting within online news documents. These baselines encompass both unsupervised sentence ranking methods and neural models.

3.6.1.1 Position-Based Baseline

News articles often adhere to the pyramid structure of reporting, with primary information predominantly contained within the initial sentences. Leveraging this structure, a position-based baseline is introduced:

- Sentences are scored in descending order based on their position from the start of the article.
- Specifically, for an article with n sentences, the i^{th} sentence is assigned a score of $1 - \frac{i}{n}$.

3.6.1.2 Graph-Based Algorithms

The study also evaluates the efficacy of renowned graph-based algorithms that exploit similarities between sentence pairs:

- **PageRank**[8]: The underlying premise of PageRank is that the significance of a webpage can be gauged by the webpages that link to it. The PageRank value for a page P is defined recursively as:

$$PR(P) = \frac{1-d}{N} + d\sum_{i=1}^{n} \frac{PR(P_i)}{L(P_i)}$$

  Where:

  - $PR(P)$ is the PageRank of page P.
  - d is a damping factor, usually set to 0.85.
  - N is the total number of pages.
  - P_i are the pages linking to page P.
  - $L(P_i)$ is the number of outbound links on page P_i.

  The algorithm involves iteratively updating the PageRank values until convergence is achieved.

- **TextRank**[46]: A popular graph-based ranking algorithm that operates on the principle of the PageRank mechanism.
- **LexRank**[21]: Similar to TextRank, LexRank also uses the PageRank algorithm but emphasizes lexical sentence similarity.

We use the position based ranking, TextRank and LexRank as our baselines

3.6.2 Evaluation Metrics

We utilized a comprehensive set of evaluation metrics to assess the performance of the models. These metrics were chosen to offer a holistic perspective on the models' capabilities, spanning from ranking accuracy to regression precision.

- **Top $\mathcal{K}$ Overlap**: This metric measures the models' accuracy in identifying the most salient sentences. Given the sets of actual and predicted top-$\mathcal{K}$ highest scored sentences, denoted as $\mathcal{A}_\mathcal{K}$ and $\mathcal{P}_\mathcal{K}$ respectively, the top $\mathcal{K}$ overlap is defined as:

$$\text{top}\mathcal{K} = \frac{|\mathcal{A}_\mathcal{K} \cap \mathcal{P}_\mathcal{K}|}{\mathcal{K}}$$

  Expressed as a percentage, it provides insights into the models' ability to correctly identify the highest scored sentences.

- **Regression Errors**: We computed the Mean Squared Error (MSE) and Mean Absolute Error (MAE) between the arrays of actual and predicted sentence labels to gauge the accuracy of the predicted scores.

- **Rank Correlation Metrics**: To understand the models' capability to rank and order the entire set of sentences in a document, We employed Spearman' s rank correlation (ρ) and Kendall' s Tau (τ). Both ρ and τ lie in the range [-1, 1].

- **nDCG**[26]: The Normalized Discounted Cumulative Gain (nDCG) metric was utilized to capture the normalized gain or usefulness of a sentence based on both its position in the inferred rank list and its actual score. nDCG values range between [0, 1].

3.6.3 Training Details

3.6.3.1 $Base_{Reg}$ Model

The $Base_{Reg}$ model was trained using the Adam optimizer[31] with a batch size of 256 documents and an initial learning rate of 10^{-5}. Default hyperparameters were employed, the maximum sequence limit for the bi-GRU layer was set to 100 sentences, and the CNN-sentence encoder had an input limit of 100 tokens per sentence.

3.6.3.2 $BERT_{Reg}$ Model

The $BERT_{Reg}$ model utilized the 6-layer bert-base-uncased variant with a maximum sequence length of 1536 tokens. The Adam optimizer was employed with a learning rate of 2.10^{-3}, and other hyperparameters were set to their default values.

3.6.4 Handling Long Documents

For documents exceeding the maximum sequence length, both the $Base_{Reg}$ and $BERT_{Reg}$ models employed a sliding window mechanism. Specifically, documents were split into overlapping sliding windows with a maximum stride of 10 sentences. During inference, the same splitting technique was adopted. For sentences within the stride of two consecutive windows, the sentence score was computed as the mean of the scores from both windows.

3.6.5 Hardware Details

All models were trained on NVIDIA RTX 2080Ti GPUs. The $Base_{Reg}$ models were trained on a single GPU for a maximum of 4 epochs with early stopping. A single epoch on the transfer learning task took over 6 hours, while one epoch for popularity forecasting training took close to 45 minutes. The $BERT_{Reg}$ models were trained on 4 GPUs for 50,000 optimizer steps with early stopping turned off for both the popularity forecasting and the transfer learning tasks, and the training time ranged between 10 to 12 hours.

3.7 Results and Discussion

Method	**TL**	**Top1**	**Top2**	**Top3**	**MSE**	**MAE**	τ	ρ	**nDCG**
Position	–	6.92	10.81	16.46	0.0079	0.0530	0.0334	0.0424	0.5804
TextRank	–	6.08	12.58	19.08	0.0316	0.1486	0.0345	0.0474	0.6313
LexRank	–	18.76	29.95	37.84	0.0072	0.0503	0.0545	0.0705	0.7324
BaseReg	X	9.12	15.49	20.93	0.0083	0.0452	0.0534	0.0746	0.6228
BaseReg	S1	9.35	14.91	20.49	0.0075	0.0478	0.0428	0.0592	0.6323
BaseReg	S2	10.08	16.63	22.83	0.0073	0.0468	0.0545	0.0751	0.6465
BaseReg	SL	9.16	14.85	20.76	0.0075	0.0475	0.0465	0.0638	0.6307
BERTReg	X	27.53	38.89	45.89	0.0055	0.0335	0.0704	0.0955	0.7921
BERTReg	S1	27.54	38.73	45.86	0.0052	0.0342	0.0734	0.0988	0.8009
BERTReg	S2	28.34	39.08	46.17	0.0053	0.0332	0.0646	0.0876	0.8009
BERTReg	SL	27.54	38.73	45.86	0.0053	0.0331	0.0510	0.0674	0.8025

Table 3.2 Consolidated results from the experiments on sentence-specific popularity forecasting. The methods are evaluated based on various metrics including Top1, Top2, Top3, Mean Squared Error (MSE), Mean Absolute Error (MAE), Spearman' s rank correlation (ρ), and Normalized Discounted Cumulative Gain (nDCG). The TL column indicates the specific Transfer Learning setup used.

3.7.1 Evaluation of Sentence Ranking Techniques

Our exploration into unsupervised sentence ranking techniques revealed distinct performance variations. LexRank[21], in particular, demonstrated superior efficacy in forecasting sentence-

specific popularity, consistently outperforming other unsupervised methods such as Position and TextRank[46] across all evaluation criteria.

3.7.2 Insights from Supervised Models

Our custom-designed supervised models, as detailed in Table 2 and 3, showcased promising results. BERTReg, in particular, emerged as the most potent architecture for predicting sentence popularity. The introduction of Transfer Learning (TL) techniques further amplified the model's performance. While each TL subtask exhibited its unique strengths, the overarching trend was clear: the integration of transfer learning invariably led to enhanced results. For instance, when BERTReg was augmented with the SL variant of TL, we observed a substantial uptick in the average nDCG, surpassing the base BERTReg model and significantly outstripping the best unsupervised methods. Statistical analyses further corroborated these findings, highlighting the robustness of BERTReg combined with TL = SL.

Interestingly, while BaseReg's standalone performance for popularity forecasting was modest, the positive influence of transfer learning was still discernible, underscoring the universal benefits of this approach.

3.7.3 Underlying Factors for Model Enhancement

The pronounced improvements observed in our models, especially with the integration of transfer learning, can be traced back to a couple of pivotal factors:

1. **Domain Consistency:** Both the popularity forecasting and salience prediction tasks drew data from online news articles. This overlap in data sources meant that transfer learning could immerse the model in a more expansive and domain-relevant dataset.

2. **Shared Task Characteristics:** Despite their distinct objectives, popularity forecasting and salience prediction share underlying similarities. Both tasks, for instance, might devalue sentences that are lexically thin or discern that certain phrases lack impactful information. By leveraging transfer learning, our models were better equipped to recognize and capitalize on these shared traits.

3.7.4 Popularity versus Salience: A Comparative Analysis

3.7.4.1 Defining Salience and Popularity

Salient sentences are those that encapsulate ideas central to the core semantics of an article, often deemed worthy of inclusion in a summary. In contrast, a sentence can deviate significantly from an article's primary topic and still be considered popular. For instance, a sentence might

not be central to an article's main theme but could contain information that resonates with a broad audience, making it popular.

3.7.4.2 Quantitative Insights

Task	Method	Top1	Top2	Top3	MSE	MAE	ρ	τ	nDCG
S1	Position	12.86	26.30	34.04	0.0010	0.0215	0.2030	0.2855	0.8682
	TextRank	15.55	24.28	30.89	0.0243	0.1487	0.0771	0.1082	0.8999
	LexRank	13.14	21.91	28.45	0.0007	0.0178	0.0561	0.0798	0.8740
	BaseReg	17.01	26.60	34.00	0.0006	0.0167	0.1387	0.1967	0.8933
	BERTReg	26.41	36.64	43.20	0.0004	0.0130	0.1372	0.1891	0.9274
S2	Position	11.24	25.03	32.77	0.0044	0.0418	0.1496	0.2101	0.7113
	TextRank	9.29	17.59	23.43	0.0280	0.1541	0.0407	0.0578	0.6665
	LexRank	11.74	20.68	26.88	0.0045	0.0422	0.0473	0.0669	0.6847
	BaseReg	17.19	27.56	35.83	0.0041	0.0373	0.1391	0.1989	0.7382
	BERTReg	23.32	36.26	43.68	0.0034	0.0332	0.1108	0.1559	0.7946
S1	Position	13.60	27.57	35.18	0.0009	0.0211	0.2050	0.2881	0.8760
	TextRank	15.55	9.29	11.72	0.0243	0.1487	0.0771	0.1082	0.8999
	LexRank	12.40	21.71	27.95	0.0007	0.0182	0.0546	0.0778	0.8657
	BaseReg	15.13	24.75	32.36	0.0007	0.0175	0.1385	0.1966	0.8780
	BERTReg	24.24	34.96	41.76	0.0005	0.0141	0.1329	0.1847	0.9152

Table 3.3 Performance of unsupervised sentence ranking baselines and proposed methods on sentence-specific popularity forecasting.

A comparative evaluation of unsupervised baselines and supervised neural models on the transfer learning subtasks revealed interesting patterns. While certain baselines, like the position-based approach, exhibited strong performance in capturing the pyramid structure of news reports, models like BERTReg excelled in metrics such as nDCG and in pinpointing the most salient sentences.

Interestingly, sentence ranking baselines appeared more adept at capturing information salience than forecasting information popularity. This distinction was evident when comparing the performance of the position baseline for both tasks. The results indicated that while the initial sentences in news articles are typically more salient, they might not always be the most popular.

3.7.4.3 Empirical Observations

Train	Eval	Top1	Top2	Top3	MSE	MAE	ρ	nDCG
S1	PF	10.97	18.31	25.69	0.0068	0.0475	0.0430	0.6864
S2	PF	10.74	19.83	27.89	0.0077	0.0455	0.0380	0.6942
SL	PF	11.44	19.09	27.02	0.0068	0.0476	0.0428	0.6937
PF	S1	8.36	16.63	22.64	0.0020	0.0301	0.0600	0.8603
PF	S2	8.51	16.34	22.53	0.0053	0.0430	0.0373	0.6579
PF	SL	9.07	16.62	22.86	0.0020	0.0304	0.0563	0.8524

Table 3.4 Cross-task evaluation - performance of BERTReg trained for popularity forecasting (PF) evaluated on salience prediction and vice-versa.

An empirical cross-task evaluation further underscored the distinction between information popularity and salience. For instance, a sentence from an article might not be deemed salient enough for summary inclusion due to its tangential relation to the article's main topic. However, it could still encompass one of the most popular information pieces within the document.

3.7.4.4 Concluding Remarks

The results and observations from this analysis highlight the nuanced differences between popularity and salience. While both concepts are integral to understanding the dynamics of online news content, they serve distinct roles and are influenced by different factors. The ability to discern between the two can offer valuable insights for various applications, from content promotion to targeted summarization.

3.8 Conclusion and Future Work

This research ventured into the novel domain of proactively forecasting sentence-specific information popularity within online news articles. The introduction of the InfoPop dataset, encompassing a vast collection of news articles labeled with normalized popularity scores, laid the foundation for our experiments. Our exploration spanned both unsupervised and supervised methodologies, with the latter benefiting significantly from a STILTs-based Transfer Learning approach rooted in salience prediction.

A key takeaway from our findings is the intricate relationship between popularity forecasting and salience prediction. While they address distinct challenges, the transfer of learning from salience prediction markedly enhanced the proficiency of our models in forecasting popularity.

This synergy underscores the potential of harnessing shared characteristics between seemingly disparate tasks.

Looking ahead, there's a rich avenue for potential applications of sentence popularity forecasting. This includes innovations in pull quote extraction and the development of popularity-guided text summarization techniques. Furthermore, a multi-task learning approach that concurrently addresses popularity forecasting and salience prediction offers an exciting direction for future research.

Chapter 4

Image-Text Multimodal Summarization

In the digital age, the vast expanse of content available on the internet has underscored the importance of effective summarization techniques. As we navigate this sea of information, the challenge is not just about condensing text but also about ensuring that the essence and context of the original content are retained. This becomes even more pertinent when we consider the inherently multimodal nature of many online articles, especially news pieces and blog posts, which seamlessly blend text with visual elements.

Our motivation to embark on this research journey stemmed from a simple observation: while textual content provides the narrative, images often capture the emotion, context, and nuances that words might miss. Could the integration of these images into the summarization process lead to richer, more comprehensive summaries?

To explore this hypothesis, we introduced "MMSumm: Multimodal Summarization of News Articles via Semantic Reranking and Cross-Modal Knowledge Distillation." This work proposes a novel knowledge-distillation based approach that aims to extract high-quality textual content. By leveraging contrastive learning with pretrained multimodal models, we sought to bridge the gap between textual and visual data, enhancing the quality and depth of generated summaries.

This chapter delves into the intricacies of our approach, shedding light on the methodologies employed, the challenges faced, and the promising results that underscore the potential of multimodal summarization in the modern information landscape.

4.1 Introduction

In the vast digital ecosystem, the art and science of summarization have become indispensable. As information continues to grow exponentially, the ability to distill this deluge into concise, meaningful summaries is not just a luxury but a necessity. Summarization, in its essence, serves as a bridge between vast information sources and the end-users, ensuring that the core message is conveyed without overwhelming the reader.

Text summarization is a pivotal computational technique designed to distill extensive textual documents into concise, coherent versions that encapsulate the primary information of the original content. The process can be bifurcated into two primary methodologies: extractive and abstractive. Extractive summarization operates by pinpointing and selecting salient segments from the source document, effectively "extracting" these portions to construct the summary. In contrast, abstractive summarization delves deeper, generating entirely new text. This approach often harnesses sophisticated language models to produce sentences that capture the overarching ideas, mirroring the synthesis one might expect from a human summarizer. As the digital realm continues to be inundated with vast amounts of information, the significance of efficient and accurate text summarization techniques becomes increasingly paramount, driving continuous advancements and refinements in the field.

Extractive summarization is a distinct approach within text summarization. It operates by selecting and "extracting" pertinent sentences or segments directly from the source document to compose the summary. This method is fundamentally data-driven, eschewing the generation of new sentences in favor of using existing content from the original document. The direct extraction offers several advantages. Firstly, it minimizes the risk of inaccuracies or misinterpretations, ensuring a high degree of accuracy. Secondly, compared to abstractive techniques, extractive methods are often more straightforward to implement, sidestepping the complexities of language generation. Lastly, the content, being directly lifted, ensures that the summary remains aligned with the tone and style of the original document. However, this approach is not without its limitations. There's potential for repetitive information if the source document contains overlapping data, leading to redundancy. Extracted sentences, when combined, might lack a natural flow, potentially yielding a disjointed summary, indicating a lack of cohesiveness. Moreover, being bound by the original text's constraints, the method might overlook nuances or broader themes that abstractive techniques could encapsulate, showcasing its limited flexibility. In the dynamic domain of text summarization, the merits and limitations of extractive methods must be judiciously weighed, especially when determining their suitability for specific summarization tasks.

Abstractive summarization represents a more sophisticated approach within the realm of text summarization. Unlike its extractive counterpart, which directly lifts segments from the source, abstractive methods generate entirely new sentences to convey the core ideas of the original content. This approach often leverages advanced language models and algorithms to craft summaries that can provide fresh perspectives or rephrased insights, mirroring human-like synthesis of information.

The primary advantage of abstractive summarization is its *flexibility*. It's not bound by the phrasing or structure of the source document, allowing for the generation of concise and often more readable summaries. This flexibility also enables the method to capture overarching themes or nuances that might be dispersed throughout the source, presenting them in a cohe-

sive manner. Moreover, abstractive methods can *reduce redundancy*, as they can synthesize information from multiple repetitive segments into a singular, coherent statement.

However, this approach comes with its set of challenges. The most notable disadvantage is the potential for *inaccuracies*. Since the method generates new content, there's a risk of introducing errors or misinterpretations that weren't present in the original document. Additionally, the complexity of abstractive algorithms often means they require more computational resources and fine-tuning, making them *computationally intensive*. Lastly, ensuring the generated content remains faithful to the original's tone and intent can be challenging, leading to potential issues with *consistency*.

In the broader landscape of text summarization, abstractive methods offer a promising avenue for capturing the essence of content in novel ways. However, their implementation requires careful consideration of their strengths and potential pitfalls.

In the realm of digital content, text-image multimodality—the integration of textual and visual data—has emerged as a powerful approach to convey and process information. Text, with its narrative strength, provides detailed descriptions, while images capture context, emotion, and intricate details that might be less effectively conveyed through words alone. The synergy of these modalities offers a comprehensive understanding, making content both engaging and memorable. This combined approach is especially beneficial in computational tasks, such as image captioning or visual question answering[62], where dual-modal input can lead to more nuanced and accurate results. Moreover, the pairing of text and images enhances accessibility, ensuring inclusivity for individuals with disabilities, and reduces potential ambiguities in content interpretation. Whether in research papers enriched with diagrams or news articles complemented by photographs, text-image multimodality offers versatility in content presentation. As digital content consumption continues to evolve, the significance of this integrated approach is set to grow, highlighting its transformative potential in both user experience and computational applications.

Multimodal summarization is an emerging frontier in the domain of text summarization, aiming to integrate multiple modalities, primarily text and visuals, to produce enriched summaries. Recognizing that information can be conveyed through various channels, this approach seeks to harness the complementary strengths of different data types to offer a more holistic understanding of the original content.

The primary advantage of multimodal summarization is its ability to provide a *richer context*. While text can convey detailed narratives, visuals often capture emotions, settings, and nuances that might be challenging to express through words alone. By integrating both, summaries can resonate more deeply with readers, offering a comprehensive insight. Additionally, in contexts where visuals play a pivotal role, such as news articles or scientific reports with crucial diagrams, multimodal summarization ensures that no critical information is lost, leading to *enhanced information retention*.

With the rise of multimodal content, especially in news articles and blogs, there's an increasing need to incorporate both text and visuals in summarization. Zhu et al.'s introduction of the Multimodal Summarization with Multimodal Output (MSMO) dataset marked a significant step in this direction. This dataset pairs text articles with associated images, providing a foundation for multimodal summarization research. However, existing models, including those based on the MSMO dataset, have limitations in effectively integrating text and visuals.

Our research is motivated by the potential benefits of multimodal content in enhancing information retention and comprehension. Recent advancements in pretrained models, such as OSCAR[36] and CLIP[54], which are trained on both text and images, offer new possibilities. By leveraging these models and incorporating techniques like cross-attentions and contrastive learning, we aim to develop a more effective multimodal summarization model. This model seeks to bridge the gap between text and visuals, producing summaries that effectively capture the essence of multimodal content.

4.2 Related Works

Text summarization, a very important task in natural language processing, aims to distill extensive documents into concise summaries that retain the core information. Two primary strategies have been prominent in this domain: extractive and abstractive summarization. Extractive methods select salient sentences or segments directly from the source [24]. On the other hand, abstractive methods, leveraging models like BERT [18] and BART [34], generate entirely new sentences to encapsulate the main ideas. BERT, introduced by Devlin et al., utilizes transformer architectures to understand deep bidirectional representations from unlabeled text. BART, a variant of BERT, focuses on denoising sequence-to-sequence pre-training, proving effective for both generation and comprehension tasks. Another notable model is BertSumm[41], an extractive summarization method that extends BERT by using interval segment embeddings. Dou et al.'s GSum [19] introduces a general framework for guided neural abstractive summarization, allowing external guidance in the form of keywords, questions, or other cues to shape the generated summary.

The rise of multimedia content has expanded the horizons of summarization to integrate multiple modalities, primarily text and visuals. Zhu et al.'s MSMO dataset [77] marked a significant step in this direction, pairing text articles with associated images. Further advancements in pretrained multimodal models, such as Oscar [36] and CLIP[54], have propelled the field. Oscar, or Object-Semantics Aligned Pre-training, aligns images and their associated textual descriptions during pre-training, enabling effective downstream vision-language tasks. CLIP (Contrastive Language–Image Pre-training) learns visual concepts from natural language supervision and has been influential in various visual tasks using the same model without task-specific tuning.

The synergy between text and images in these models offers a comprehensive context, ensuring summaries that are both detailed and engaging.

4.3 Problem Statement

Given an article $\mathcal{A}$ containing t tokens and a set of images $\mathcal{I}$, the objective is to generate a summary $\mathcal{A}'$ containing t' tokens and an image $\mathcal{I}'$ such that $t' < t$ and $\mathcal{I}' \in \mathcal{I}$ is the most relevant to both $\mathcal{A}$ and $\mathcal{A}'$. The challenge is further compounded by the absence of ground truth image labels for the training set.

The proposed model, as illustrated in Fig. 1 of the paper, comprises three primary modules: an encoder, a multimodal scorer, and a decoder. The encoder is responsible for independently encoding the input text and images. The multimodal scorer evaluates the relevance of the images in relation to the text, and the decoder generates the final summary by integrating the textual and visual information.

The overarching goal is to ensure that the generated summary not only captures the essence of the article but also incorporates context derived from the most pertinent image, resulting in a comprehensive and enriched summary.

4.4 Dataset Description

The primary dataset employed for this research is the *MSMO (Multimodal Summarization with Multimodal Output)* dataset. Below are the key characteristics and details of this dataset:

- **Origin**: The MSMO dataset is curated from articles sourced from the CNN and Daily Mail websites[77].
- **Composition**: Each article in the dataset is paired with multiple images. Specifically, the median number of images associated with each article is 6.
- **Size**: The dataset encompasses a total of 314,581 articles, cumulatively containing over 1.5 million images.
- **Selection for Study**: Due to computational and space constraints, only the top 7 images from each article were considered for this study. The selection was based on the images' order of occurrence within the article.
- **Reference Images**: While ground truth images are not provided for the training phase, a list of reference images is included for evaluation purposes.

- **Enhancements**: Zhu et al. further refined the dataset by introducing a golden reference image for each data point. This was achieved using ranking techniques, such as rouge-based overlap and order of occurrence in the document.

This rich multimodal dataset serves as the foundation for training and evaluating the proposed summarization model, allowing for a comprehensive assessment of the model's ability to generate summaries that integrate both textual and visual information.

4.5 Architecture

The problem statement we are trying to solve can be formally described as follows - Given an article containing text T and a set of images I, we need to generate a textual summary t such that $length(t) < length(T)$. Along with the summary we also need to return the most relevant image $i \in I$ that complements the generated summary. In the dataset, while we possess ground truth summaries for training the textual part of the model, we do **not** have the access to ground truth relevant images while training (however, these are present in the validation/testing iterations).

In the sections that follow, we describe the modules within are architecture in detail. Fig. 6.1 shows an overview of the model.

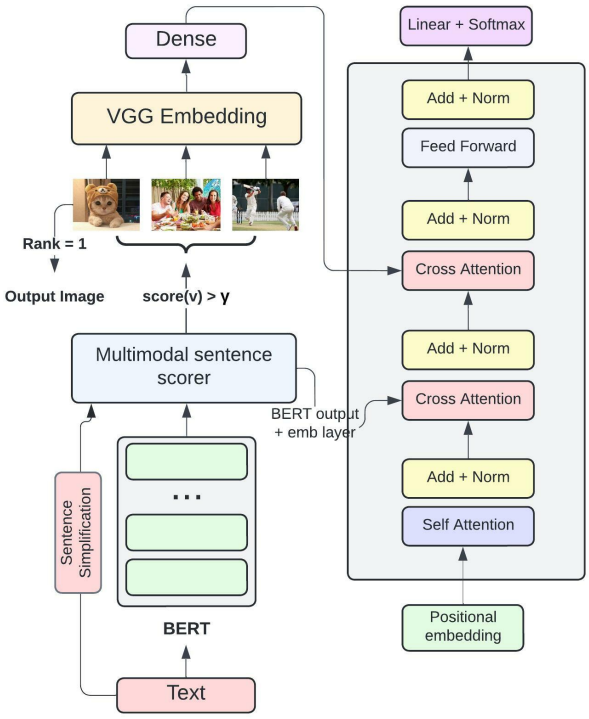

Figure 4.1 Model Overview

4.5.1 Sentence Simplification Module

The Sentence Simplification module is an integral part of the proposed architecture, designed to preprocess and refine the input text to ensure compatibility with the subsequent stages of the model, especially the OSCAR model.

4.5.1.1 Objective

The primary objective of this module is to simplify the sentences by removing specific named entities and aligning the text with the characteristics of the pretrained models. This alignment is crucial because the OSCAR mode;, which plays a significant role in the architecture, is pretrained on the COCO dataset, known for its simple object annotations.

4.5.1.2 Named Entity Replacement

To achieve the desired simplification, named entities within the sentences are identified and replaced. The replacement process leverages GloVe[51] word embeddings to find the closest COCO class corresponding to each named entity. This ensures that the text is in a format that resonates well with the OSCAR model's training data.

4.5.1.3 Removal of Abstract Objects

Beyond named entities, the module also identifies and removes phrases containing abstract objects, such as indications of time or place. The Spacy library[24], a popular tool for natural language processing, aids in this removal process. By eliminating these abstract references, the module further streamlines the text, ensuring clarity and compatibility with the model's subsequent stages.

4.5.1.4 Significance

The Sentence Simplification module plays a pivotal role in enhancing the model's output quality. By refining the input text, it ensures that the multimodal scorer can effectively evaluate the relevance of images in relation to the text, leading to more coherent and contextually relevant summaries.

4.5.2 Encoder

The encoder plays a pivotal role in the proposed model, serving as the initial step in processing both textual and visual inputs. It comprises two distinct components, each tailored to handle a specific modality:

4.5.2.1 Text Encoder

The Text Encoder is responsible for transforming the input text from the article into a suitable representation. This representation captures the semantic essence of the text, ensuring that the core information is retained for subsequent processing. The encoder leverages deep learning architectures, likely transformers or recurrent neural networks, to encode the sequence of words into a fixed-size vector representation. This representation serves as the foundation for the subsequent steps in the summarization process.

Specifically, the model employs *BERT* [18] to generate contextual embeddings of size $N \times D$, where N represents the number of tokens and D is the hidden dimension of the architecture. Given that BERT has a maximum context size of 512 tokens, documents are truncated at this limit to fit within the constraints.

4.5.2.2 Image Encoder

The Image Encoder, on the other hand, processes the associated images from the article. Its primary objective is to encode the visual content into a format that can be seamlessly integrated with the textual data. Given the complexity and richness of visual data, the encoder employs convolutional neural networks (CNNs) or other advanced vision models to extract salient features from the images. These features capture the visual semantics, ensuring that the most pertinent visual information is retained for the summary generation.

The model utilizes *VGG-19* [60] embeddings, generated from the 'fc2' layer of the deep convolutional network. A threshold θ is set, and embeddings of the top-θ images, as determined by the multimodal scorer, are considered. The size of the returned embeddings is $\theta \times D'$. If a document contains fewer than θ images, the embeddings are padded with zero vectors. To ensure compatibility with the cross-attention mechanism within the architecture, a linear transformation is applied to these embeddings.

In essence, the encoder architecture ensures that both textual and visual modalities are processed optimally, setting the stage for the generation of comprehensive and enriched summaries that seamlessly integrate information from both sources.

4.5.3 Multimodal Scorer

The Multimodal Scorer serves as a crucial component in the proposed architecture, designed to assess the semantic similarity between text and images. This scoring mechanism ensures that the most contextually relevant visual information is integrated with the textual content for the generation of the summary.

4.5.3.1 Objective

The primary aim of this module is to bridge the gap between the textual and visual modalities. By evaluating the semantic similarity between text and images, the scorer aids in determining the relevance of each image to the associated text.

4.5.3.2 Scoring Mechanism

For each article, the scorer calculates the semantic similarity between every pair of text and image. This is achieved by feeding the model with text and the corresponding image features (extracted using faster-rcnn[56] for the OSCAR model). The output is a scalar value that indicates the semantic similarity of the image to the text. Formally, for each article A with a set $S = \{s_1, s_2, ..., s_n\}$ of sentences and a set $I = \{i_1, i_2, ..., i_m\}$ of images, the score $\sigma(s, i)$ is defined for every pair of text s and image i.

$$\sigma_{avg}(s) = \frac{1}{m} \sum_{j=1}^{m} \sigma(s, i_j) \tag{4.1}$$

$$\sigma_{avg}(i) = \frac{1}{n} \sum_{j=1}^{n} \sigma(s_j, i) \tag{4.2}$$

Images are then ranked based on $\sigma_{avg}(i)$. Those with a rank beyond a threshold are discarded, ensuring only the top-ranked images are considered for the summary.

4.5.3.3 Utilized Models

Two state-of-the-art multimodal models, OSCAR[36] and CLIP[54], are employed within this module. These models are adept at understanding visual-linguistic tasks, ensuring a robust evaluation of text-image pairs.

4.5.3.4 Significance

The Multimodal Scorer is instrumental in ensuring that the generated summaries are both coherent and contextually rich. By selecting the most relevant visual content, it adds depth and context to the summaries, making them more comprehensive.

4.5.4 Image Selection Process

The image selection process is pivotal in the multimodal summarization model, aiming to identify the most pertinent image that aligns with the textual content of the summary. The selection is driven by a multimodal scorer, which computes a relevance score for each image in relation to the textual content.

The score for each sentence s and image i is defined as:

$$S(s) = \sum_{i=1}^{N} f(s, i)$$

$$S(i) = \sum_{s=1}^{M} f(s, i)$$

Where f represents the function that computes the relevance between a sentence and an image, N is the total number of images, and M is the total number of sentences.

Images are then ranked based on their scores $S(i)$. Images with a rank greater than a predefined threshold τ are discarded. The top-τ images, based on their relevance scores, are retained for the decoding process.

This selection mechanism ensures that the chosen image is not only congruent with the textual content but also augments the overall understanding and appeal of the summary.

4.5.5 Cross-Attention and Decoder

The Cross-Attention and Decoder module is integral to the proposed architecture, ensuring the effective fusion of textual and visual modalities during the summary generation process.

4.5.5.1 Cross-Attention Mechanism

The cross-attention mechanism is designed to facilitate the interaction between the encoded textual and visual representations. This mechanism allows the model to weigh the importance of different parts of the text and image, ensuring that the most relevant information from both modalities is considered during the decoding process. Specifically, the model employs an extra cross-attention layer in the decoder, which takes the transformed VGG-embeddings of images and the BERT embeddings of text as input. This cross-attention allows the model to focus on specific regions of the image and corresponding parts of the text, enhancing the coherence and relevance of the generated summary.

4.5.5.2 Decoder Architecture

The decoder is responsible for generating the final summary based on the processed textual and visual inputs. After the cross-attention layer, the remaining decoder layers adhere to the standard transformer layout. The decoding strategy employed by the model is beam search with multiple beams, ensuring a diverse set of candidate summaries and selecting the most appropriate one.

4.5.5.3 Significance

The Cross-Attention and Decoder module is instrumental in ensuring that the generated summaries are both coherent and contextually rich. By effectively integrating information from both textual and visual modalities, it produces summaries that are not only informative but also engaging, capturing the essence of the original content.

4.5.6 Contrastive Learning in Multimodal Summarization

Contrastive learning, in the context of multimodal summarization, plays a pivotal role in enhancing the quality of generated summaries by effectively leveraging both textual and visual information. The proposed model in the paper utilizes contrastive learning to guide text generation with image inputs and vice versa.

4.5.6.1 Principle and Application

The model employs a contrastive objective to improve the Rouge[37] scores of the generated summaries. By using contrastive loss, the model shifts its focus on including tokens that are ranked highly by the multimodal scorer. This ensures that the tokens related to the images and those in the gold summary are closely aligned, leading to summaries that are semantically closer to the gold standard.

4.6 Training Methodology

The training process of the proposed multimodal summarization model is meticulously designed to ensure the effective integration of both textual and visual information. The methodology encompasses various components, including the loss functions, optimization strategies, and specific hyperparameters.

4.6.1 Loss Functions

The primary loss function employed during training is the Negative Log Likelihood (NLL) loss, defined as:

$$L_{\text{NLL}} = -\frac{1}{|\mathcal{T}|} \sum_{t=1}^{|\mathcal{T}|} \log p(t|\mathcal{C}), \tag{4.3}$$

where $p(t|\mathcal{C})$ represents the probability of token t given the previous tokens and context $\mathcal{C}$.

To induce an image correspondence to the loss, a contrastive objective based on the OSCAR scores (S) and the parameter θ is introduced:

$$L_{\text{contrastive}} = -\frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}}\log\left(\frac{1-S(s)}{S(s)+|\mathcal{T}|}\right) \tag{4.4}$$

where $\mathcal{S}$ represents the set of tokens for which $S > \theta$ and $\mathcal{T}$ represents the set of tokens for which $S \leq \theta$.

The final loss is a combination of the NLL loss and the contrastive loss:

$$L = \lambda L_{\text{NLL}} + (1-\lambda)L_{\text{contrastive}}, \tag{4.5}$$

where λ is a hyperparameter.

4.6.2 Optimization and Hyperparameters

The model employs the Adam[31] optimizer for gradient updates. The encoder utilizes BERT weights from a BertSum checkpoint, and these weights are frozen during training. The decoder consists of 6 layers with 8 attention heads per layer. A learning rate decay of 0.02 is applied after 8000 warmup steps. The decoding strategy is beam search with 5 beams. The base setup of OSCAR[36] and the ViT-B-32 version of CLIP[54] are used for the multimodal scorers.

4.6.3 Implementation Details

The text is encoded using BERT-base, resulting in a hidden dimension of 768. For the multimodal sentence scorer, the parameters are set as $d = 3$ and $\theta = 0.01$. The value of λ in the final loss is set to 0.7. The model is trained on an Nvidia GTX 1080 Ti, taking approximately one day for the entire process.

4.7 Experiments

4.7.1 Baselines

To evaluate the effectiveness of our proposed model, we compare it against the following baselines:

1. **MSMO**: This model employs a bidirectional LSTM[23] for text embedding and VGG-19[60] for image embedding. An attention mechanism is utilized to integrate both modalities, and the resultant output is decoded using a unidirectional LSTM layer.

2. **BertSum**[41]: This model leverages a fine-tuned BERT for text encoding and a conventional transformer decoder for summary generation.

3. **UniMS**[76]: A multimodal summarization model that adopts BART's[34] encoder-decoder architecture to facilitate information transfer between modalities.

These baselines offer a comprehensive comparison across diverse architectures and methodologies. While the MSMO model signifies a traditional approach using LSTMs and CNNs, BertSum acts as a potent text-only baseline harnessing BERT's capabilities, and UniMS represents a contemporary multimodal approach built on BART's[34] architecture.

The code for MSMO and UniMS is not publically available, so we compare our findings with their reported ROUGE scores only and Image Precision calculated over identical train-val-test splits.

4.7.2 Evaluation Metrics

To rigorously assess the performance of the proposed model, the following evaluation metrics were employed:

- **ROUGE (R1, R2, RL)**[37]: A standard metric in summarization tasks, ROUGE measures the overlap between the n-grams in the generated summary and the reference summary. The metric provides insights into the precision, recall, and F1 score of the summaries. The ROUGE scores are given by:

$$\text{ROUGE-N} = \frac{\sum_{s \in \text{ref}} \sum_{\text{gram}_n \in s} \text{Count}_{\text{match}}(\text{gram}_n)}{\sum_{s \in \text{ref}} \sum_{\text{gram}_n \in s} \text{Count}(\text{gram}_n)}$$

  where $\text{Count}_{\text{match}}$ is the maximum number of times a gram appears in both the reference and the candidate summary.

- **BERTScore (BERT-F1)**[75]: This metric leverages contextualized BERT embeddings to calculate similarity. It offers an evaluation on a semantic level. BERTScore is computed as the cosine similarity between BERT embeddings of the generated and reference summaries.

- **Image Precision (IP)**: Given the multimodal nature of the task, Image Precision was introduced to evaluate the relevance of the selected images in the summaries. The IP is given by:

$$IP = \frac{\text{ref}_{\text{img}} \cap \text{rec}_{\text{img}}}{\text{rec}_{\text{img}}}$$

  where ref_{img} and rec_{img} refer to the reference images and model-recommended images, respectively.

- **Human Scores**: Apart from automated metrics, human evaluation was conducted on the outputs of the model. Annotators were asked to rate randomly sampled predicted summaries and associated images on a scale of 1-5, with 1 being incomprehensible and 5 being human-like.

These metrics collectively offer a comprehensive evaluation of the model's performance, considering both textual and visual aspects of the generated summaries.

4.7.3 Results

The evaluation of our proposed model's performance was conducted against several baselines using a diverse set of evaluation metrics. Our model demonstrated a significant improvement in ROUGE scores when the contrastive objective was incorporated. This enhancement in BERTScore[75] indicated that the summaries generated were semantically more aligned with the gold standard. Additionally, the increase in Human scores suggested that the summaries were more satisfactory to human evaluators. In terms of Image Precision (IP), the model's performance was commendable, showcasing that the images selected by the model were highly relevant and closely matched the reference images. Notably, our model, especially the variant employing the contrastive loss, surpassed the performance of a robust text-only baseline, BertSum[41]. This superiority underscores the benefits of integrating multimodal information into the summarization process, leading to the generation of high-quality multimodal summaries that are both semantically precise and visually pertinent.

Model	*R1*	*R2*	*RL*	*BERT-F1 Scores*	*Human Scores*	*Image Precision*
MSMO	40.86	18.27	37.75	-	-	62.44
BertSum	42.13	19.60	39.18	0.810	2.80	-
UniMS	**42.94**	**20.50**	**40.96**	-	-	69.38
Ours$_{CLIP}$	42.28	19.37	39.81	0.852	3.01	67.41
Ours$_{OSCAR}$	42.51	19.97	39.28	**0.882**	**3.17**	**77.99**

Table 4.1 Comparison of our models with the baselines.

4.8 Discussions

4.8.1 Discussion

The introduction of contrastive learning in our model emerged as a pivotal factor in enhancing the quality of generated summaries. This approach aimed to ensure that the model focuses on tokens that are ranked highly by our multimodal scorer, thereby emphasizing the correlation between tokens related to the images and those present in the gold summary.

The use of contrastive loss shifted the model's attention towards creating summaries that not only align with the textual content but also resonate with the visual context provided by the images. This was evident from the significant improvement in ROUGE[37] scores and BERTScore when the contrastive objective was utilized. The increase in these scores indicated

that the generated summaries were more semantically aligned with the gold standard and effectively incorporated visual cues.

Furthermore, our findings based on the results showcased that the model variant with the contrastive loss outperformed other variants, solidifying the importance of this loss function in the task of multimodal summarization. The contrastive loss ensured that the model's focus was directed towards tokens that were of higher relevance, leading to the generation of summaries that were both semantically precise and visually congruent.

In conclusion, the incorporation of contrastive learning in our model has proven to be a significant advancement in the realm of multimodal summarization, emphasizing the synergy between textual and visual modalities to produce high-quality summaries.

4.9 Conclusion

4.9.1 Conclusion and Future Works

In this work, we explored the potential of leveraging multimodal information for the task of summarization. Our proposed architecture effectively utilized both textual and visual cues to produce high-quality summaries. The results underscored the viability of our approach, especially when contrastive learning was incorporated, leading to summaries that were both semantically aligned with the gold standard and visually congruent.

The societal implications of our work are profound. By increasing the accessibility of crucial news through platforms like social media, we can potentially reduce the spread of misinformation. Summarization models, when guided by well-curated training data rather than personal biases or agendas, inherently produce more neutral and less biased content. This is crucial in today's digital age, where the rapid dissemination of information is paramount.

Looking ahead, there are several avenues for future research. We aim to further enhance our architecture by employing more robust text and image encoders. These improvements, although computationally intensive, hold the promise of pushing the boundaries of multimodal summarization. Additionally, the exploration of other modalities and their integration into the summarization process could pave the way for even more comprehensive and informative summaries.

Chapter 5

Salience Guided Summarization

This chapter delves into an innovative approach to text summarization by integrating a nuanced sentence salience scaling technique, initially developed in a prior work, MMSUMM, into the renowned GSum model. The exploration pivots on modulating sentence encodings based on their salience scores, aiming to refine the abstractive summarization process by ensuring that the most pertinent sentences predominantly influence the generated summaries. While the model does not directly incorporate multimodal data, the conceptual framework derived from MMSUMM provides a unique lens through which the summarization process is viewed and manipulated. Despite the model not yielding the anticipated satisfactory results, this exploration unveils critical insights and learning points, offering a valuable foundation upon which future research can build, particularly in the realm of effectively utilizing sentence salience in the text summarization process.

5.1 Introduction

Text summarization, a pivotal domain within Natural Language Processing (NLP), has witnessed significant advancements, yet the journey towards generating succinct, coherent, and contextually rich summaries continues to present intriguing challenges and opportunities. The essence of summarization lies in its ability to distill voluminous textual data into concise representations, thereby facilitating enhanced information retrieval, comprehension, and utility for end-users.

The motivation behind the work presented in this chapter emanates from the desire to further refine the summarization process, ensuring that generated summaries are not only succinct but also deeply rooted in the most salient aspects of the original text. The GSum model[19], renowned for its efficacy in extractive summarization, serves as a foundational pillar in our exploration. GSum adeptly identifies and extracts the most salient sentences from a document, which subsequently guide the abstractive summarization process. However, the model, while

proficient, does not inherently modulate the influence of extracted sentences based on their varying degrees of salience within the abstractive summarization phase.

In light of this, our work seeks to integrate a concept previously developed in a model named MMSUMM, which employs a technique of scaling sentence encodings based on their salience scores. The primary advantage of this approach lies in its potential to further refine the abstractive summarization process, ensuring that sentences of higher salience exert proportionally greater influence on the generated summary. This nuanced approach aims to enhance the relevance and focus of the resultant summaries, ensuring they are tightly aligned with the most critical aspects of the original text.

The objective of this chapter is twofold: firstly, to explore the viability and efficacy of integrating salience-based encoding scaling into the GSum model; and secondly, to evaluate the impact of this integration on the quality and relevance of the generated summaries. While our model does not directly incorporate multimodal data, the conceptual underpinning derived from MMSUMM, which was originally developed in a multimodal context, provides a unique and innovative approach to enhancing unimodal text summarization.

As we navigate through the intricacies of this adaptation, we shall explore the challenges encountered, the insights gleaned, and the potential pathways that future research might explore in the continual pursuit of advancing the field of text summarization. Despite the results of this exploration not aligning with initial expectations, the learnings derived therein provide a valuable stepping stone for future endeavors in the realm of text summarization, particularly in leveraging sentence salience to enhance summary generation.

5.2 Model Architecture

5.2.1 Extractive Summarizer

The Extractive Summarizer, denoted as E, is a pre-trained model that identifies salient sentences S from the input document D. It assigns a salience score s_i to each sentence i in D. For this work we use BertSumm_Ext[41] as our extractive summarizer

$$S = E(D)$$

5.2.2 Dual Encoder

The Dual Encoder consists of two separate encoders: the Document Encoder E_D and the Retrieved Sentences Encoder E_S.

5.2.2.1 Document Encoder

The Document Encoder E_D processes D and scales the sentence encodings using the salience scores s_i from E.

$$E_{D_i} = E_D(D_i) \cdot s_i, \quad \forall i \in D$$

5.2.2.2 Retrieved Sentences Encoder

The Retrieved Sentences Encoder E_S processes the salient sentences S without scaling.

$$E_{S_i} = E_S(S_i), \quad \forall i \in S$$

5.2.3 Decoder

The Decoder D_C utilizes two cross attention layers to process the encodings from E_D and E_S, generating the summary Y.

$$Y = D_C(E_{D_i}, E_{S_i}), \quad \forall i \in D, S$$

The first cross attention layer processes E_{D_i}, and the second processes E_{S_i}. The output summary Y is generated by attending to both sets of encodings sequentially.

5.3 Experiments

5.3.1 Objective

The primary objective of our experiments is to outperform the GSum model[19] in text summarization, exploring innovative adaptations and methodologies. We focus on enhancing the summarization quality by integrating an encoding scaling layer, which modulates sentence encodings based on their salience, into the GSum framework. Our experiments are bifurcated into two primary settings: Oracle and non-Oracle, each providing unique insights into the model's performance and potential areas for improvement.

5.3.2 Dataset

For this experiment we utilise the CNN/DM dataset[47] for text summarization. The CNN/Daily Mail (CNN/DM) dataset is one of the most widely used datasets for evaluating text summarization models. It is derived from online news articles from CNN and the Daily Mail. The dataset is known for its substantial size and has been utilized in numerous studies as a benchmark for both extractive and abstractive summarization tasks.

The CNN/DM dataset comprises over 280,000 training instances, 13,000 validation instances, and 11,000 test instances, making it one of the largest summarization datasets available. Each instance in the dataset consists of a news article and a summary. The summaries in the CNN portion are derived from bullet points (highlights) that accompany the articles, providing a succinct overview of the main news points. Meanwhile, the Daily Mail summaries are obtained from labeled sentences within the associated articles. This dataset poses various challenges for summarization models due to its real-world, noisy web text, and the diversity of topics covered, making it a robust choice for evaluating the generalization and effectiveness of different summarization approaches.

5.3.3 Experiment Setup

In the Oracle setting, we utilize pre-annotated gold summaries as the extractive summary input for the abstractive summarization phase. The salience scores for sentences are binary, determined by whether a sentence is present in the gold summary.

The GSum_adapted_Oracle model integrates an encoding scaling layer into the GSum framework, aiming to enhance the abstractive summarization by scaling sentence encodings with their respective binary salience scores.

The Oracle setting serves as an upper-bound performance benchmark, providing insights into the best possible summarization outcomes when optimal extractive summaries are utilized. Comparing GSum_Oracle and GSum_adapted_Oracle allows us to evaluate the efficacy of the encoding scaling layer when the extractive summarization is optimal.

In the non-Oracle setting, we employ BertSumm[41], an extractive summarizer, to generate the salient sentences and their respective salience scores, which are then used in the abstractive summarization phase.

5.3.4 Model Variants: GSum+BertSumm and GSum_adapted+BertSumm

- **GSum+BertSumm:** Utilizes BertSumm for extractive summarization, feeding its output into the original GSum model for abstractive summarization. - **GSum_adapted+BertSumm:** Integrates the encoding scaling layer into GSum, modulating the sentence encodings with salience scores derived from BertSumm during the abstractive summarization phase.

5.3.5 Metrics

For this experiment we employed ROUGE[37]. The ROUGE (Recall-Oriented Understudy for Gisting Evaluation) framework is a standard evaluation metric for assessing the quality of text summaries. It quantitatively measures the similarity between system-generated summaries and reference summaries, typically crafted by humans, by analyzing their n-gram overlap.

5.3.6 Key Variants

- **ROUGE-N:** Evaluates n-gram overlap, with common variants being ROUGE-1 (unigram) and ROUGE-2 (bigram).
- **ROUGE-L:** Considers the longest common subsequence (LCS), capturing the largest co-occurring sequence of words, irrespective of word order.

5.4 Results

Model	*R1*	*R2*	*RL*
GSUM_Oracle	**55.18**	**32.54**	**52.06**
GSUM_adapted_Oracle	51.73	28.27	48.07
GSUM+BertSumm	43.78	20.66	40.66
GSUM_adapted+BertSumm	41.36	18.42	38.48

Table 5.1 Comparison of our models with the baselines.

The models were evaluated based on their ROUGE[37] scores (R1, R2, and RL), as presented in Table 1. GSum_Oracle, utilizing pre-annotated gold summaries, achieved the highest scores across all three ROUGE metrics, serving as a robust baseline. The adapted models, despite their innovative encoding scaling layer, did not surpass the baseline, indicating potential areas for further refinement and exploration in future work.

5.5 Conclusion

We conducted an experiment to adapt the GSum[19] work while utilising core principles for our work on multimodal summarization shown in the previous chapter. Unfortunately a direct adaption our work on guided summarization did not lead to improvements over the GSum model. This indicates that either the concepts used in multimodal summarization does not translate to improvements in guided summarization or that there is a need for further refinement of the scaling layer for this task.

Chapter *6*

Persona Based Summarization

The challenge of distilling vast and intricate legal documents into concise summaries is a pressing concern in the digital age, especially when catering to diverse stakeholders. "Aspect-based Summarization of Legal Case Files using Sentence Classification" addresses this very challenge, focusing on the complexities inherent to SEBI case files. The study underscores the criticality of providing stakeholders with streamlined access to essential information, eliminating the need to navigate through dense legal jargon. By targeting the unique nuances of Indian legal adjudicating orders, the research highlights the broader implications of effective summarization in the legal domain. The overarching goal is to harness advanced computational techniques to produce summaries that are not only concise but also tailored to the specific needs of varied personas, such as Investors and Defense Lawyers. This endeavor epitomizes the broader ambition of the text summarization field: to transform voluminous data into actionable insights, ensuring that vital information is both accessible and comprehensible to all.

6.1 Introduction

In today's rapidly digitizing world, the legal sector stands at the crossroads of tradition and innovation. With an ever-growing corpus of legal documents, there's an increasing demand to make this information more digestible and accessible to a broader audience. The paper titled "Aspect-based Summarization of Legal Case Files using Sentence Classification" delves deep into this challenge, specifically targeting the intricate and often convoluted SEBI case files.

The Securities and Exchange Board of India (SEBI) plays a pivotal role in regulating India's securities and commodity market. As a result, a plethora of legal documents, ranging from regulations to adjudications, are generated under its purview. For stakeholders, ranging from corporate entities to individual investors, navigating this maze of information can be daunting. The need for concise, clear, and contextually relevant summaries has never been more pressing.

However, the task is far from straightforward. Legal documents are characterized by their dense terminology, intricate structures, and domain-specific nuances. Moreover, different stake-

holders require different facets of information. For instance, while an investor might be keen on understanding the implications of a particular adjudication, a defense lawyer would be more interested in the nuances of the arguments presented.

This research aims to bridge this gap. By harnessing the power of computational techniques and diving deep into the realm of SEBI case files, the study seeks to develop methodologies that can produce tailored summaries, catering to the unique needs of diverse stakeholders. In doing so, it hopes to set a precedent for how legal document summarization can be approached in the digital age, ensuring that vital legal information is not just available but also easily comprehensible.

6.2 Related Works

The evolution of text summarization, particularly within the legal domain, has been marked by a series of innovative methodologies and techniques aimed at condensing extensive information into succinct and informative summaries.

Aspect-based summarization has emerged as a prominent approach, with early works focusing on the summarization of product reviews by emphasizing specific product attributes, such as a car's durability. This methodology has been expanded to encapsulate specific topics within expansive documents.

In the realm of legal text processing, the adaptation of BERT[18] for legal contexts, notably LEGAL-BERT by Chalkidis et al. LegalBERT[10], has set a benchmark for the nuanced processing of legal texts. This adaptation ensures contextually relevant interpretations and processing of legal documents.

Another significant stride in the field is the LexRank algorithm by Erkan and Radev[21], which employs graph-based lexical centrality. This approach is adept at pinpointing salient information within texts, proving invaluable for gleaning key insights from verbose legal documents.

Furthermore, the application of Convolutional Neural Networks (CNN) for sentence classification, as proposed by Yoon Kim, offers a robust mechanism for segmenting and categorizing sentences within extensive texts. This technique has demonstrated efficacy in differentiating diverse informational types within a document.

Building upon these foundational works, the present research endeavors to advance the frontiers of legal text summarization, with a specific focus on SEBI case files.

6.3 Dataset

The pursuit of effective aspect-based summarization in the legal domain necessitates a dataset that is both comprehensive and tailored to capture the nuances of legal language. To

address this need, a dataset was curated by extracting over 7000 adjudication orders from the SEBI website. Of these, 27 adjudication orders were meticulously annotated with the expertise of a legal professional, yielding 2264 distinct sentence-label pairs that served as the foundation for model training.

These adjudication orders were specifically centered around the Prohibitions of Insider Trading (PIT) regulations. This focus ensures a dataset that is deeply rooted in the complexities of insider trading within the Indian legal context. This dataset stands out as a pioneering effort in the realm of Indian legal adjudication orders on insider trading, providing a valuable resource for future research.

The process of label conceptualization was grounded in a detailed preliminary study of a random selection of case files. This rigorous analysis led to the identification of distinct classes of sentences that are typically present in such legal documents. Based on this study, a set of labels was conceptualized:

- **Material Fact**: Pertinent information that forms the crux of the case.
- **Procedural Fact**: Details about the procedural aspects of the case.
- **Statutory Fact**: Information related to specific statutes or laws.
- **Related Fact**: Information that provides context or background.
- **Issue Framed**: The primary issues or questions raised in the case.
- **Subjective Observation**: Observations made based on personal judgment.
- **Defendant Claim**: Statements highlighting the defendant's stance.
- **Allegation**: Accusations or charges levied in the case.
- **Penalty**: Details about any penalties imposed or discussed.
- **Violation**: Statements discussing specific regulatory violations.

To ensure the reliability and consistency of the annotations, a subset of 10 documents (approximately 40% of the data) was re-annotated by a second legal expert. This step was crucial to measure inter-rater reliability, ensuring that the labels were consistently applied and capturing the inherent subjectivity of legal document annotation.

In essence, this dataset not only lays the groundwork for the current research but also offers a structured approach for aspect-based summarization, enabling the extraction of specific facets of information tailored to diverse stakeholder needs.

6.4 Model Description

The task of aspect-based summarization of legal case files, especially those as intricate as SEBI adjudication orders, demands a sophisticated model architecture that can capture the nuances and intricacies of legal language. Our approach is a multi-step process, encompassing several modules designed to work in tandem.

6.4.1 Sentence Classification Module

The foundation of our approach is the semantic segmentation of the legal text into sentences, each associated with one of ten predefined legal labels. For this, we employed a fine-tuned version of the uncased BERT-Base model, which underwent training for 70,000 steps on approximately 7000 SEBI adjudication orders. The tasks included Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). To enhance the context-awareness of our model, we incorporated a window of context at the sentence level. Within this window, ten sentences were chosen and processed using the hierarchical attention mechanism, as proposed by Yang et al. [72]. This mechanism weighs the sentences based on their importance, producing a single context vector for both the left and right context segments. This vector, when concatenated with the target sentence's embedding, is then passed through a multi-layer perceptron to generate the final predictions.

6.4.2 Aspect-Based Filtering Module

Post-classification, the next step is to filter sentences relevant to the respective stakeholders based on the labels generated. The stakeholders and their interests are as follows:

- **Investors**: Primarily interested in Material Facts, which provide the core narrative of the adjudication proceedings, details about company shareholdings, and the alleged violations. Penalties offer insights into the consequences of the determined violations.
- **Defense Lawyers**: In addition to Material Facts, they focus on Defendant Claims, which shed light on the defendant's perspective, and Issues Framed, which highlight the central disputes within the case.
- **Adjudicating Officers**: Apart from Material Facts and Penalties, they value Related Facts, which offer reasoning behind final orders and help in linking similar case files.

6.4.3 Summarization Module

For the actual summarization, we leveraged the BRIO abstractive summarization model[42], which is trained on the CNN-Daily Mail dataset[47]. This model employs a transformer-based

encoder-decoder framework, introducing noise to its inputs during training and learning to reconstruct them. To ensure that the summaries retained most of the input information, the stakeholder-specific input was divided into chunks of three sentences. These chunks were then processed by the model, and the outputs were concatenated to produce the final aspect-based summary.

In essence, our model architecture is a harmonious blend of classification, filtering, and summarization, tailored to meet the unique demands of legal document summarization.

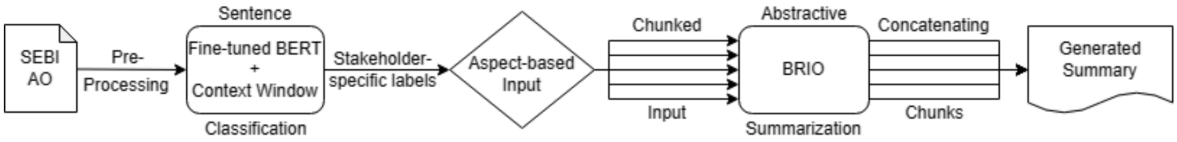

Figure 6.1 Aspect-Based Summarization Pipeline

6.5 Experiments

6.5.1 Sentence Classification

6.5.1.1 Baselines

The intricate nature of legal language and the precision required in classifying sentences within legal documents necessitate a robust evaluation framework. To benchmark the performance of our sentence classification approach, we compared it against a spectrum of models, ranging from classical machine learning techniques to advanced neural architectures.

6.5.1.1.1 Classical Machine Learning Models In the realm of classical machine learning, several models were paired with diverse embedding techniques to capture the semantic nuances of the sentences. Models such as **Logistic Regression**[16], **Random Forest**[7], **SVM**[15], and **XGBoost**[11] were experimented with. These models were complemented with embeddings derived from techniques like **Word2Vec**, **tf-idf**, **Fine-tuned ELMO**, and **GloVe**[51]. The combination of these models and embeddings aimed to provide a holistic evaluation of the sentence classification task.

6.5.1.1.2 Classical Neural Models Neural architectures, with their ability to discern intricate patterns in data, were also explored. The **CNNs with trainable Word2Vec embeddings** model[30] was employed, harnessing the power of convolutional layers to process sentence embeddings. Additionally, the **BiLSTM with attention**[39] model was utilized, leveraging

GloVe embeddings as input. This model's attention mechanism ensures that the most salient parts of the sentences are emphasized during classification.

6.5.1.1.3 Transformer-based Models The transformative capabilities of transformer architectures in natural language processing are undeniable. In this context, models like **XL-Net**[71], **Uncased BERT-Base**, and **LEGAL-BERT**[10] were considered. These transformer-based classifiers, with their self-attention mechanisms, offer a nuanced approach to sentence classification, especially in the domain of legal texts.

Together, these baselines provide a comprehensive landscape against which our sentence classification approach was evaluated, ensuring a rigorous assessment of its capabilities.

6.5.1.2 Metrics

The **F1 score** is a harmonic mean of precision and recall, providing a balance between the two. It is particularly useful when class distributions are imbalanced. The F1 score is defined as:

$$F1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

Where *precision* is the ratio of correctly predicted positive observations to the total predicted positives, and *recall* is the ratio of correctly predicted positive observations to the all observations in actual class.

6.5.2 Results

The sentence classification task aimed to semantically segment legal case files into predefined classes. To evaluate the performance of our approach, we utilized a dataset comprising 27 annotated documents, resulting in 2264 sentence-label pairs. The performance metrics for both neural and classical machine learning methods were considered.

Our model, which incorporated context by considering sentences before and after the target sentence, showcased promising results. A notable observation was the varying performance across different labels. For instance, sentences labeled as 'Procedural Facts' and 'Subjective Observations' frequently appeared as clusters within a document, leading to higher accuracy in their classification. On the other hand, labels like 'Penalty' posed challenges, possibly due to their sparse occurrence in the dataset.

A deeper dive into label-wise metrics revealed that 'Penalty' and 'Procedural Fact' labels exhibited the lowest and highest performances, respectively. The inclusion of context significantly improved the classification accuracy for sentences adjacent to those with the same label. For example, sentences grouped under 'Procedural Facts' and 'Subjective Observations' predominantly appeared in clusters within documents, leading to more accurate classifications.

Embeddings→	Word2Vec			tf-idf			Finetuned ELMO			GloVe		
Classifier↓	*Acc*	$F1_m$	$F1_w$	*Acc*	$F1_m$	$F1_w$	*Acc*	$F1_m$	$F1_w$	*Acc*	$F1_m$	$F1_w$
LR	0.647	0.599	0.637	0.664	0.564	0.654	**0.706**	**0.657**	**0.699**	0.653	0.515	0.631
RF	0.660	0.622	0.660	0.671	0.671	0.671	0.656	0.595	0.648	0.578	0.550	0.577
SVM	0.675	0.607	0.665	0.679	0.613	0.669	0.648	0.593	0.635	0.653	0.555	0.643
XGBoost	0.651	0.619	0.651	0.678	0.659	0.678	0.649	0.605	0.639	0.571	0.547	0.571

Table 6.1 Classical ML Method Results for Sentence Classification

Table 1 in the paper provides a comprehensive breakdown of the performance metrics for various machine learning methods and embeddings. Classical machine learning methods, when paired with embeddings like Word2Vec, tf-idf, Fine-tuned ELMO[52], and GloVe[51], demonstrated varying degrees of success. The results underscore the importance of choosing the right combination of model and embeddings for optimal performance in sentence classification within legal documents.

Model	*Accuracy*	$F1_{macro}$	$F1_{weighted}$
Fine-tuned BERT + Context Window	**83.56**	**0.795**	**0.830**
Fine-tuned BERT + Two Sided Context	78.06	0.750	0.780
BERT	73.46	0.680	0.730
Legal BERT	70.91	0.670	0.710
XLNet	73.29	0.620	0.700
BiLSTM + Attention	64.12	0.570	0.630
CNN-non-static	68.00	0.650	0.680

Table 6.2 Neural Method Results for Sentence Classification

In conclusion, the sentence classification module effectively segments legal case files, acting as a semantic layer that aids legal experts in quickly assimilating pertinent information.

6.5.3 Text Summarization

6.5.3.1 Baselines

The task of summarizing legal documents, with their inherent complexity and precision, necessitates a rigorous evaluation framework. To this end, our approach was benchmarked against a range of established methodologies, both extractive and abstractive in nature.

6.5.3.1.1 Unsupervised Extractive Models Extractive summarization techniques focus on identifying and extracting pivotal sentences or segments directly from the source text. Among the unsupervised extractive models employed, **TextRank**[46] stands out as an adaptation of the PageRank algorithm, tailored for text, ranking sentences based on their significance.

In a similar vein, **LexRank**[21] determines the importance of sentences using the concept of eigenvector centrality in a graph representation. Additionally, the **STAS** approach [70] leverages the self-attention weights of sentences, emphasizing the most salient ones in the summary.

6.5.3.1.2 Abstractive Models Diverging from extraction, abstractive summarization seeks to generate novel sentences that encapsulate the essence of the original content. The **BERTSum**[41]model, with its encoder-decoder architecture, utilizes BERT[18] as its encoder, making it adept at crafting abstractive summaries. Another notable model is **BART** [34], a transformer-based architecture trained to reconstruct texts by introducing noise during the training phase, further enhancing its abstractive capabilities.

In sum, these baselines offer a comprehensive backdrop for evaluating the performance of our summarization technique, ensuring its robust assessment within the domain of legal document summarization.

6.5.3.2 Metrics

To ensure a comprehensive evaluation of our methodologies for both sentence classification and summarization, a blend of intrinsic and extrinsic metrics was employed. These metrics provide insights into the quality, coherence, and relevance of the generated outputs in comparison to the source documents.

6.5.3.2.1 Intrinsic Metrics for Summarization Intrinsic metrics offer a deep dive into the inherent qualities of the generated outputs:

Semantic Similarity (SS): This metric measures the cosine similarity between the source document (D) and the generated summary (S).

$$SS(D, S) = \frac{D \cdot S}{\|D\|\|S\|}$$

Compression (C): Representing the ratio of the number of tokens in the summary (n_S) to those in the input document (n_D).

$$C = 1 - \frac{n_S}{n_D}$$

Redundancy (R): By computing the average cosine similarity between all sentence pairs within a generated summary.

$$R = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1, j \neq i}^{N} \text{cosine\_similarity}(S_i, S_j)$$

Where N is the number of sentences in the summary.

Coherence (CH): Evaluating the fluency of the generated summary, coherence is determined by computing the likelihood of each sentence occurring after its predecessor.

6.5.3.2.2 Extrinsic Metrics for Summarization Extrinsic metrics provide an external evaluation:

ROUGE Score[37]: A widely-used metric in the realm of text summarization, ROUGE evaluates text similarity by calculating token overlaps between the generated summary and a reference summary.

BERTScore[75]: Going beyond token overlaps, BERTScore evaluates similarity at a semantic level.

In conclusion, the combination of these intrinsic and extrinsic metrics offers a holistic evaluation framework, ensuring that our methodologies are both accurate and effective in the context of legal document processing.

6.6 Results

Our summarization experiments were meticulously conducted on a dataset comprising 1000 adjudication orders, which were previously labeled by the sentence classification module. To further refine the evaluation, a specialized chunking pipeline was integrated with the top-performing method, facilitating its assessment on the gold summaries derived from 27 adjudication orders. Despite the inherent challenges posed by the scarcity of legal data, the primary objective was to gauge the generalization capabilities of the models to our dataset.

6.6.1 Intrinsic Metrics Evaluation

Model	Semantic Similarity	Compression	Redundancy	Coherence
TextRank	0.765	0.583	0.257	0.125
LexRank	0.812	0.542	0.234	0.145
STAS	0.781	0.610	0.213	0.120
BERTSum$_{ExtAbs}$	0.776	0.878	0.193	0.327
BART	0.815	0.853	**0.154**	0.412
BRIO	**0.827**	**0.894**	0.178	**0.481**

Table 6.3 Intrinsic Metrics for Summarization

The intrinsic metrics, designed to delve into the inherent attributes of the generated summaries, painted a vivid picture of each model's performance. **TextRank**[46], for instance, exhibited a commendable semantic similarity of 0.765, but its coherence at 0.125 indicated room for improvement. On the other hand, **LexRank**[21] showcased a slightly superior semantic similarity of 0.812, with its coherence also marginally better at 0.145.

STAS[70] demonstrated a balanced performance with a semantic similarity of 0.781 and coherence of 0.120. Diving into the realm of abstractive methods, **BERTSum**[41] achieved a

semantic similarity of 0.776 and an impressive compression of 0.878. **BART**[34] further elevated the benchmarks with a semantic similarity of 0.815 and a coherence of 0.412. However, the best performing model was **BRIO**[42], which outshone its counterparts with a stellar semantic similarity of 0.827 and coherence of 0.481.

The overarching inference drawn was the relative ineffectiveness of extractive methods, which often retained non-essential parts, leading to heightened redundancy. In stark contrast, abstractive techniques, epitomized by BRIO, emerged as the frontrunners, crafting summaries that were both concise and coherent.

6.6.2 Extrinsic Metrics Evaluation

Model	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
BART	0.416	0.261	0.352	0.792
BRIO	**0.454**	**0.271**	**0.359**	**0.809**

Table 6.4 Extrinsic Metrics for Summarization with Chunked Input

The extrinsic metrics, tailored to offer a comparative evaluation against established benchmarks, further elucidated the prowess of the models. **BART**, for instance, secured ROUGE-1, ROUGE-2, and ROUGE-L scores of 0.416, 0.261, and 0.352, respectively, complemented by a BERTScore[75] of 0.792. Yet, it was **BRIO** that truly shined, registering ROUGE[37] scores of 0.454, 0.271, and 0.359 for ROUGE-1, ROUGE-2, and ROUGE-L, respectively, and a BERTScore of 0.809.

6.6.2.1 Aspect-based Summarization

A significant facet of our summarization experiments was the generation of persona-specific summaries. The objective was to cater to the distinct informational needs of various stakeholders, namely the Investor, Adjudicating Officer, and Defense Lawyer. The results, as presented in Table 5 of the paper, highlighted the differential performance across these personas.

Aspect	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
Adjudicating Officer	0.455	0.262	0.330	0.803
Defence Lawyer	0.424	0.241	0.320	0.798
Investor	**0.484**	**0.310**	**0.395**	**0.826**

Table 6.5 Persona-based Metrics for Chunked Input using BRIO

For the **Adjudicating Officer**, the ROUGE-1, ROUGE-2, and ROUGE-L scores were 0.455, 0.262, and 0.330, respectively, with a BERTScore of 0.803. The **Defense Lawyer** summaries yielded scores of 0.424, 0.241, and 0.320 for ROUGE-1, ROUGE-2, and ROUGE-L, respectively, complemented by a BERTScore of 0.798. The **Investor** summaries emerged as the best-performing, with ROUGE scores of 0.484, 0.310, and 0.395, and a BERTScore of 0.826.

The superior performance of Investor summaries can be attributed to their relatively non-technical nature. Investors primarily seek material facts and final penalties, devoid of intricate legal nuances. In contrast, Defense Lawyers and Adjudicating Officers require a deeper dive into the legal intricacies, encompassing issues framed and associated facts.

The intrinsic evaluation showcased the relative strengths and weaknesses of the employed models. Abstractive methods, especially BRIO, emerged as the frontrunners, crafting summaries that were both concise and coherent. The inability of certain models to adapt well to the legal domain was attributed to their training on non-legal datasets. To mitigate this, aspect-based filtering using sentence classification was employed, enhancing the relevance and accuracy of the generated summaries.

In conclusion, the summarization experiments underscored the efficacy of our approach in generating persona-specific summaries. The results highlight the potential of aspect-based summarization in catering to the diverse informational needs of various stakeholders in the legal domain.

6.7 Conclusion and Future Work

In our endeavor, we have introduced a robust system tailored to generate personalized summaries of intricate legal case files, aligning with the unique requirements of different stakeholders. This system not only facilitates the generation of summaries but also semantically segments legal case files into well-defined categories. Such segmentation acts as a pivotal semantic layer, empowering legal professionals to swiftly assimilate and comprehend crucial information.

The creation of a novel dataset, encompassing Indian legal documents, stands as a testament to our commitment to advancing the domain of legal document analysis. This dataset, we believe, will serve as a valuable resource for future explorations and analyses of Indian legal documents.

While our system has showcased promising results, the realm of legal document summarization is vast and ever-evolving. Future endeavors could delve into length-controlled abstractive summarization techniques[57, 40], which possess the potential to adapt the encoding of input based on the desired summary length. Such advancements could further refine the quality and relevance of generated summaries, ensuring they cater even more effectively to the diverse needs of legal stakeholders.

Chapter 7

Indic Multimodal Data Creation

In the rapidly evolving landscape of machine learning, multimodal models, exemplified by innovations like DALL-E, have emerged as versatile tools with applications extending beyond academia. However, a significant challenge remains: the scarcity of clean, large-scale data, especially for languages other than English. Recognizing this gap, this chapter underscores the pressing need for multimodal datasets tailored for the rich tapestry of Indian languages. Motivated by the goal of democratizing multimodal machine learning in the Indian context, the research proposes the creation of a comprehensive Image-text pair multimodal dataset encompassing 11 Indian languages. This ambitious endeavor seeks to prune the existing Samanantar dataset[55] to yield high-quality caption-like sentences, setting the stage for a transformative shift in multimodal research in India. The ultimate vision is not just to curate a dataset but to pave the way for future innovations, fostering a more inclusive and diverse multimodal research landscape.

7.1 Introduction

Machine learning models, particularly multimodal ones such as DALL-E and ImaGen, require extensive datasets for effective training and application. While there's an abundance of data in English, there's a noticeable gap when it comes to Indian languages. This research addresses this gap by focusing on the development of a multimodal dataset for 11 Indian languages. Using the Samanantar dataset as a starting point, the goal is to extract and refine data to produce a comprehensive Image-text pair dataset tailored for these languages. This effort is not just about data accumulation; it's a strategic move to enhance the quality and scope of multimodal research specific to the Indian linguistic context.

7.1.1 Related Works

7.1.1.1 Multimodal Models

- **DALL-E**: Developed by OpenAI, this model is designed to generate images from textual descriptions, showcasing the potential of integrating vision and language.
- **StableDiffusion**: Another notable model in the multimodal domain.
- **ImaGen**: A testament to the versatility of multimodal models.
- **OSCAR and CLIP**: These are vision-language pre-trained models. CLIP[54] (Contrastive Language-Image Pretraining) by OpenAI is particularly significant for its capability to jointly understand images and text, trained on a vast dataset of internet text paired with images.

7.1.1.2 Multimodal Datasets in English

- **MS-COCO**: A dataset by Microsoft, it features over 300,000 images, each accompanied by 5 captions, focusing on common objects within 81 categories.
- **Flickr8k**: Comprises 8,000 images from Flickr, each with 5 captions, chosen for their depiction of diverse scenes and situations.
- **Conceptual Captions**[58]: Introduced by Google, this dataset pairs images with more descriptive captions, offering a richer context compared to MS-COCO.

7.1.1.3 Indic Multimodal Datasets

- **Hindi Visual Genome Dataset**: A pioneering dataset for English-Hindi multimodal machine translation. It consists of English segments from the Visual Genome, paired with images. These segments are translated into Hindi, with manual post-editing considering the associated images.
- **Samanantar Dataset**[55]: A text-only dataset from AI4Bharat, it contains 49.7 Million pairs of sentences between English and 11 Indic languages, spanning the Indo-Aryan and Dravidian language families. The dataset serves as a foundation for creating a comprehensive Image-text pair multimodal dataset for these languages.

7.1.2 Motivation

A predominant challenge is the absence of multimodal models tailored for non-English text. The majority of advanced multimodal models have been designed and trained with a focus

on English-centric datasets. This approach has inadvertently led to an underrepresentation of non-English languages, especially those from linguistically diverse regions like India. The implications of this oversight are manifold, limiting the potential applications and benefits of multimodal research in non-English speaking regions.

Further compounding this challenge is the lack of authentic multimodal datasets. While there are datasets available, such as the "Hindi Visual Genome", they are primarily constructed from translated content. These translated datasets, albeit valuable, often lack the authenticity and nuances inherent to native language data. The consequence is models that may overlook the rich cultural and contextual subtleties, which are pivotal for accurate representation and understanding.

Lastly, there's a pressing need to expand the horizons of multimodal summarization to encompass non-English languages. The capabilities of multimodal summarization hold immense promise, with the potential to redefine content consumption patterns in non-English languages. Through this research, by developing genuine multimodal datasets and models for languages like Hindi and other Indic languages, there's an aspiration to extend our work on multimodal summarization to a broader demographic. This endeavor is not just about technological advancement; it's about ensuring that content is both contextually relevant and culturally resonant, democratizing access to cutting-edge machine learning models.

7.2 Goal

The main aim of this work is to create a text-image pair dataset encompassing 11 Indian Languages. Since text captions and images are the most related pairs of text and image we try to build a dataset of images and captions. Unfortunately such a dataset does not exist and to create such a dataset a lot of annotation effort is required along with the requirement of a large number of annotators as our dataset contains text from 11 languages. So we come up with a solution that requires very little manual intervention to build such a dataset. The main use of this dataset is to use this dataset to train CLIP[54] like visuolinguistic transformer models that can be used for various tasks like caption generation, multimodal summarization for non English languages

7.3 Dataset Creation Methodology

The foundation for our dataset creation in Indic languages is the *Samanantar* dataset, which boasts 42 million sentence pairs spanning English and various Indic languages.

1. **Rule-based Pruning**: The process commences with a rule-based approach, filtering out sentences with structures that are not apt for captions, such as those containing abstract nouns.

2. **Caption Classifier**: A dedicated classifier is trained to pinpoint potential English captions. This step is further enhanced using the *text-davinci-003* model from the GPT-3[**?**] family for refinement.

3. **Image Retrieval**: For each shortlisted caption, the top three corresponding images are sourced using Google Image Search. The CLIP[54] model is then employed to ascertain the semantic congruence between the image and the text.

4. **Post-processing**: The final phase involves refining captions. This includes simplifying certain named entities and excising temporal or locative details.

This streamlined methodology is geared towards curating an authentic Indic dataset, setting the stage for a more granular exploration in the ensuing sections.

7.4 Grammatical Rule-Based Pruning

In the endeavor to curate high-quality captions, it's imperative to ensure that the textual descriptions adhere to certain grammatical standards. These standards not only ensure the clarity and coherence of the captions but also facilitate better learning for visuolinguistic transformers. The following are the defined features for a potential caption:

1. **No Imperatives**: Captions should be descriptive and not instructive. Hence, imperative forms of verbs, which often indicate commands or requests, are avoided.

2. **Action Verbs**: A good caption often describes an action. Therefore, the presence of action verbs is encouraged. However, the tense of these verbs is restricted to either the present or past participle to maintain consistency and clarity.

3. **Third Person Perspective**: Captions should maintain an objective tone, and thus, they are written in the third person. This ensures that the description remains neutral and universally applicable.

4. **Single Clause**: To maintain simplicity and directness, captions are restricted to a single clause. This ensures that the description is concise and directly related to the image.

5. **Simple Noun Phrases**: For captions that primarily describe objects, simple noun phrases are preferred. This ensures that the focus remains on the object without introducing unnecessary complexity.

6. **Statements Over Questions**: Captions should provide information and not seek it. Therefore, they are framed as statements rather than questions.

In the dataset creation process, the initial step involves rule-based pruning, where sentences are filtered based on the aforementioned grammatical features. We do this using Spacy's NLP parsers[24] that allow us to prune out sentences based on grammatical features that dont adhere to our idea of a caption. This approach ensures that the curated captions are not only grammatically sound but also contextually relevant to the images they describe. By adhering to these rules, the dataset aims to maintain a high standard of quality, ensuring that the captions are both informative and aligned with the visual content.

7.5 Caption Classifier

The objective of creating a caption classifier is to accurately and efficiently identify sentences that can serve as potential captions for images. Given the vastness of the *Samanantar* dataset[55], an automated approach is essential to sift through the data and pinpoint relevant captions.

7.5.1 Aims

The classifier is designed to:

- Distinguish between generic sentences and those that can act as descriptive captions for images.
- Ensure high precision, focusing on minimizing false positives to maintain the quality of the dataset.

7.5.2 Training Dataset Creation

To train the caption classifier, a balanced dataset is curated:

- **Positive Samples**: These are sourced from established English image-text datasets, ensuring that they are genuine captions. Datasets like MSCOCO, Flickr8k, and ConceptualCaptions contribute to these samples.
- **Negative Samples**: These are derived from generic text datasets, such as CNN-DM[47] and Books. The aim is to include sentences that are structurally or semantically unfit to be captions.

The training set is deliberately skewed to contain a higher number of negative samples (400k) compared to positive samples (50k) to build a high precision classifier.

7.5.3 Classifier Models and Results

Various models were explored to achieve the desired precision. Traditional machine learning classifiers like CatBoost[68] and XGBoost[11] were tested alongside transformer-based classifiers such as DistilBERT and RoBERTa[43]. Some models were also fed with sentences augmented with parts-of-speech tags, considering the structural importance in determining captions.

The evaluation was conducted on a manually labeled test set of 200 sentences, categorized as either captions or non-captions. Among the tested models, certain transformer-based classifiers demonstrated promising precision, indicating their potential in accurately identifying captions.

7.5.4 Refinement with GPT-3

To further enhance the classifier's performance, the *text-davinci-003* model from the GPT-3 family[**?**] was integrated. Preliminary tests on a subset of sentences showcased promising results, suggesting its efficacy in refining the caption selection process.

In conclusion, the caption classifier is a pivotal component in the dataset creation pipeline. Through rigorous training and refinement, it ensures that the curated dataset is of high quality, containing genuine and contextually relevant captions.

7.6 Image Retrieval

The image retrieval process is a crucial step in the dataset creation pipeline, ensuring that each caption is paired with a relevant image. This process is multi-faceted, involving both automated image search and subsequent validation using advanced models.

7.6.1 Automated Image Search

For the set of sentences filtered through the caption classifier, images are sourced using an automated Google Images scraper. This scraper fetches images based on the content of each sentence, aiming to retrieve visuals that best represent the described scenario or object.

7.6.2 Image-Text Relevance Validation

Given the vast and varied nature of the internet, not all retrieved images may be contextually aligned with their corresponding captions. To address this, the CLIP (Contrastive Language–Image Pre-training) model is employed. CLIP[54] is designed to understand images paired with natural language, making it apt for this validation task.

Each image-text pair is passed through the CLIP model to gauge their semantic similarity. Pairs that don't meet a predefined similarity threshold are pruned, ensuring that the dataset contains only coherent and contextually relevant image-text combinations.

7.6.3 Mapping to Indic Translations

Once the English captions and their corresponding images are finalized, the next step is to map these captions to their Indic translations. Leveraging the *Samanantar* dataset, each English caption is paired with its translation in the target Indic language. This results in a multimodal dataset that not only represents the visual content but also caters to multiple Indian languages.

7.7 Post-Processing

After the meticulous process of image retrieval and mapping captions to their Indic translations, further refinement of the dataset is undertaken through post-processing. This step ensures that the Indic language sentences are simplified and standardized, making them more suitable for training multimodal models.

7.7.1 Simplification of Sentences

Complex sentences, especially those with multiple clauses, can introduce ambiguity and make it challenging for models to draw clear correlations between text and image. To address this, additional clauses are pruned, and sentences are streamlined to convey a singular, clear idea.

7.7.2 Named Entity Replacement

Named entities, while informative, can be overly specific and may not always align with the generic nature of the associated image. For instance, a specific person's name might be replaced with a more generic term like "man" or "woman". This step ensures that the captions remain general and can be effectively used to train models that aim for broader applicability.

7.8 Conclusion

This work successfully curated a multimodal dataset for 11 Indic languages. Through a series of systematic steps, including rule-based pruning, caption classification, image retrieval, and post-processing, a reliable dataset was constructed. The primary outcome is a dataset that pairs visual content with corresponding textual descriptions in multiple Indian languages. The next objective is to utilize this dataset to train effective multimodal models, further advancing the capabilities of machine learning models in the context of Indic languages.

Chapter *8*

Conclusion and Future Work

This thesis journeyed through various facets of text summarization, emphasizing the enhancement of the domain through multifarious approaches.

8.1 Sentence Popularity Forecasting

Diving into the domain of sentence-specific information popularity within online news articles, our research introduced the pivotal InfoPop dataset. Our foray into both unsupervised and supervised techniques, particularly the STILTs-based Transfer Learning approach, revealed the intertwined nature of popularity forecasting and salience prediction. Harnessing the shared characteristics between these tasks showcased vast potential. Future directions encompass innovations such as pull quote extraction, popularity-guided text summarization, and an integrative multi-task learning approach targeting both popularity forecasting and salience prediction.

8.2 Multimodal Summarization

Our exploration into multimodal information for summarization led to a promising architecture that harmoniously integrated textual and visual cues. By leveraging contrastive learning, we achieved summaries that demonstrated both semantic and visual alignment. The profound societal implications of this work, especially in mitigating misinformation spread, underscore its significance. As we move forward, our vision includes refining our architecture with robust encoders and exploring other modalities to further enrich multimodal summarization.

8.3 Guided Summarization

Our efforts to adapt concepts from our multimodal summarization into the realm of guided summarization, specifically building upon the GSum model, unfortunately, didn't yield the

expected enhancements. This underscores the intricacies of translating improvements across different summarization tasks and paves the way for future investigations into refining the scaling layer or other integral components for optimal results.

8.4 Legal Document Summarization

Legal document summarization presented a unique challenge, catering to diverse stakeholder needs. We introduced a comprehensive system that goes beyond mere summarization by semantically segmenting legal case files. The creation of a novel dataset for Indian legal documents further solidified our contributions. The world of legal document summarization is vast, and future endeavors could explore length-controlled abstractive summarization techniques to cater more effectively to varied legal needs.

8.5 Multimodal Dataset for Indic Languages

Lastly, our dedicated efforts resulted in the curation of a unique multimodal dataset for 11 Indic languages. This dataset, a blend of visual content and textual descriptions, promises to be an invaluable resource. With this dataset in hand, our next steps will focus on leveraging it to train advanced multimodal models, thus further enhancing machine learning capabilities for Indic languages.

8.6 Final Remarks

In summation, this thesis has journeyed through enriching text summarization from various angles - from predicting sentence popularity to delving into multimodal approaches and catering to niche domains like legal document summarization. Each stride has contributed to the broader goal of enhancing the utility, relevance, and accessibility of summaries in our ever-evolving digital age.

Related Publications

1. Sayar Ghosh Roy, **Anshul Padhi**, Risubh Jain, Manish Gupta, Vasudeva Varma **Towards Proactively Forecasting Sentence-Specific Information Popularity within Online News Documents** Proceedings of the 33rd ACM Conference on Hypertext and Social Media [HT '22] (Main Track)

2. **Anshul Padhi**, Tanmay Sachan, Balaji Vasan Srinivasan, Vasudeva Varma **MMSumm: Multimodal Summarization via Semantic Reranking and Cross-Modal Knowledge Distillation**. Publication under review.

3. **Anshul Padhi**, Nikhil E, Pulkit Parikh, Swati Kanwal, Kamal Karlapalem, Natraj Raman **Aspect-based Summarization of Legal Case Files using Sentence Classification** WWW '23 Companion: Companion Proceedings of the ACM Web Conference 2023

Bibliography

[1]

[2] R. Abidi, Y. Xu, J. Ni, W. Xiangmeng, and w. Zhang. Popularity prediction of movies: from statistical modeling to machine learning techniques. *Multimedia Tools and Applications*, 79:1–35, 12 2020.

[3] M. S. Ageev, D. Lagun, and E. Agichtein. Improving search result summaries by using searcher behavior data. *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, 2013.

[4] S. Angelidis and M. Lapata. Summarizing opinions: Aspect extraction meets sentiment prediction and they are both weakly supervised, 2018.

[5] A. Balali, M. Asadpour, and H. Faili. A supervised method to predict the popularity of news articles. *Computación y Sistemas*, 21, 01 2018.

[6] P. Bhaskar and S. Bandyopadhyay. Language independent query focused snippet generation. In *CLEF*, 2012.

[7] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[8] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1):107–117, 1998. Proceedings of the Seventh International World Wide Web Conference.

[9] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners, 2020.

[10] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos. LEGAL-BERT: the muppets straight out of law school. *CoRR*, abs/2010.02559, 2020.

[11] T. Chen and C. Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM.

[12] T. Chen, L. Wu, X. Li, J. Zhang, H. Yin, and Y. Wang. Call attention to rumors: Deep attention based recurrent neural networks for early rumor detection, 2017.

[13] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling, 2014.

[14] J. M. Conroy, J. D. Schlesinger, and D. P. O'Leary. Topic-focused multi-document summarization using an approximate oracle score. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 152–159, Sydney, Australia, July 2006. Association for Computational Linguistics.

[15] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[16] D. R. Cox. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2):215–232, 1958.

[17] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[18] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[19] Z. Dou, P. Liu, H. Hayashi, Z. Jiang, and G. Neubig. Gsum: A general framework for guided neural abstractive summarization. *CoRR*, abs/2010.08014, 2020.

[20] E. Egonmwan, V. Castelli, and M. A. Sultan. Cross-task knowledge transfer for query-based text summarization. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 72–77, Hong Kong, China, Nov. 2019. Association for Computational Linguistics.

[21] G. Erkan and D. R. Radev. Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Intell. Res. (JAIR)*, 22:457–479, 2004.

[22] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith. Don't stop pretraining: Adapt language models to domains and tasks, 2020.

[23] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[24] M. Honnibal and I. Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017.

[25] M. Hu and B. Liu. Mining and summarizing customer reviews. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004.

[26] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, Oct. 2002.

[27] C. Jiang, W. Ye, H. Xu, S. Huang, F. Huang, and S. Zhang. Vision language pre-training by contrastive learning with cross-modal similarity regulation. In *Proceedings of the 61st Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 14660–14679, Toronto, Canada, July 2023. Association for Computational Linguistics.

[28] Z. Junnan, Y. Zhou, J. Zhang, H. Li, C. Zong, and C. Li. Multimodal summarization with guidance of multimodal reference. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:9749–9756, 04 2020.

[29] Y. Keneshloo, S. Wang, E. Han, and N. Ramakrishnan. Predicting the popularity of news articles. In *SDM*, 2016.

[30] Y. Kim. Convolutional neural networks for sentence classification. *CoRR*, abs/1408.5882, 2014.

[31] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization, 2017.

[32] K. Kreimeyer, M. Foster, A. Pandey, N. Arya, G. Halford, S. F. Jones, R. Forshee, M. Walderhaug, and T. Botsis. Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. *Journal of Biomedical Informatics*, 73:14–29, 2017.

[33] K. Lerman and R. McDonald. Contrastive summarization: An experiment with consumer reviews. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 113–116, Boulder, Colorado, June 2009. Association for Computational Linguistics.

[34] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, 2019.

[35] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang. Visualbert: A simple and performant baseline for vision and language, 2019.

[36] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, Y. Choi, and J. Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. *CoRR*, abs/2004.06165, 2020.

[37] C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[38] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár. Microsoft coco: Common objects in context, 2015.

[39] G. Liu and J. Guo. Bidirectional lstm with attention mechanism and convolutional layer for text classification. *Neurocomputing*, 337, 02 2019.

[40] Y. Liu, Q. Jia, and K. Zhu. Length control in abstractive summarization by pretraining information selection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6885–6895, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[41] Y. Liu and M. Lapata. Text summarization with pretrained encoders. *CoRR*, abs/1908.08345, 2019.

[42] Y. Liu, P. Liu, D. Radev, and G. Neubig. Brio: Bringing order to abstractive summarization, 2022.

[43] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

[44] J. Lu, D. Batra, D. Parikh, and S. Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, 2019.

[45] Q. Lu, X. Ye, and C. Zhu. Mtca: A multimodal summarization model based on two-stream cross attention. In *2022 2nd International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI)*, pages 594–601, 2022.

[46] R. Mihalcea and P. Tarau. TextRank: Bringing order into texts. In *Proceedings of EMNLP-04and the 2004 Conference on Empirical Methods in Natural Language Processing*, July 2004.

[47] R. Nallapati, B. Xiang, and B. Zhou. Sequence-to-sequence rnns for text summarization. *CoRR*, abs/1602.06023, 2016.

[48] R. Nallapati, F. Zhai, and B. Zhou. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. *arXiv preprint arXiv:1611.04230*, 2016.

[49] S. Parida, O. Bojar, and S. R. Dash. Hindi visual genome: A dataset for multimodal english-to-hindi machine translation, 2019.

[50] T. Penin, H. Wang, T. Tran, and Y. Yu. Snippet generation for semantic web search engines. In *ASWC*, 2008.

[51] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[52] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. In *Proc. of NAACL*, 2018.

[53] D. Radev, T. Allison, S. Blair-Goldensohn, J. Blitzer, A. Çelebi, S. Dimitrov, E. Drabek, A. Hakim, W. Lam, D. Liu, J. Otterbacher, H. Qi, H. Saggion, S. Teufel, M. Topper, A. Winkel, and Z. Zhang. MEAD - a platform for multidocument multilingual text summarization. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May 2004. European Language Resources Association (ELRA).

[54] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision, 2021.

[55] G. Ramesh, S. Doddapaneni, A. Bheemaraj, M. Jobanputra, R. AK, A. Sharma, S. Sahoo, H. Diddee, M. J, D. Kakwani, N. Kumar, A. Pradeep, S. Nagaraj, K. Deepak, V. Raghavan, A. Kunchukuttan, P. Kumar, and M. S. Khapra. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages, 2021.

[56] S. Ren, K. He, R. B. Girshick, and J. Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *CoRR*, abs/1506.01497, 2015.

[57] I. Saito, K. Nishida, K. Nishida, A. Otsuka, H. Asano, J. Tomita, H. Shindo, and Y. Matsumoto. Length-controllable abstractive summarization by guiding with summary prototype. *CoRR*, abs/2001.07331, 2020.

[58] P. Sharma, N. Ding, S. Goodman, and R. Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*, 2018.

[59] T. Shi, Y. Keneshloo, N. Ramakrishnan, and C. K. Reddy. Neural abstractive text summarization with sequence-to-sequence models, 2020.

[60] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

[61] H. Singh. Predicting the popularity of online news using social features. *2018 Second International Conference on Green Computing and Internet of Things (ICGCIoT)*, pages 514–518, 2018.

[62] H. Singh, A. Nasery, D. Mehta, A. Agarwal, J. Lamba, and B. V. Srinivasan. MIMOQA: Multimodal input multimodal output question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5317–5332, Online, June 2021. Association for Computational Linguistics.

[63] H. V. Singh. Predicting the popularity of online news using social features. In *2018 Second International Conference on Green Computing and Internet of Things (ICGCIoT)*, pages 514–518, 2018.

[64] A. Tatar, P. Antoniadis, M. Amorim, and S. Fdida. Ranking news articles based on popularity prediction. pages 106–110, 08 2012.

[65] M. T. Uddin, M. J. A. Patwary, T. Ahsan, and M. S. Alam. Predicting the popularity of online news from content metadata. In *2016 International Conference on Innovations in Science, Engineering and Technology (ICISET)*, pages 1–5, 2016.

[66] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

[67] A. Voronov, Y. Shen, and P. K. Mondal. Forecasting popularity of news article by title analyzing with bn-lstm network. In *Proceedings of the 2019 International Conference on Data Mining and Machine Learning*, ICDMML 2019, page 19–27, New York, NY, USA, 2019. Association for Computing Machinery.

[68] K. Wang, P. Wang, X. Chen, Q. Huang, Z. Mao, and Y. Zhang. A feature generalization framework for social media popularity prediction. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 4570–4574, New York, NY, USA, 2020. Association for Computing Machinery.

[69] B. Wu, W.-H. Cheng, Y. Zhang, Q. Huang, J. Li, and T. Mei. Sequential prediction of social media popularity with deep temporal context networks, 2017.

[70] S. Xu, X. Zhang, Y. Wu, F. Wei, and M. Zhou. Unsupervised extractive summarization by pre-training hierarchical transformers. *CoRR*, abs/2010.08242, 2020.

[71] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[72] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California, June 2016. Association for Computational Linguistics.

[73] J. Zhang, Y. Zhao, M. Saleh, and P. J. Liu. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization, 2019.

[74] L. Zhang, X. Zhang, J. Pan, and F. Huang. Hierarchical cross-modality semantic correlation learning model for multimodal summarization, 2021.

[75] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. Bertscore: Evaluating text generation with bert, 2020.

[76] Z. Zhang, X. Meng, Y. Wang, X. Jiang, Q. Liu, and Z. Yang. Unims: A unified framework for multimodal summarization with knowledge distillation, 2021.

[77] J. Zhu, H. Li, T. Liu, Y. Zhou, J. Zhang, and C. Zong. MSMO: Multimodal summarization with multimodal output. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4154–4164, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics.