TEAM 4 : Language models, Classification and Summarization

CS5604: Information Storage and Retrieval

Under the guidance of Dr. Edward Fox

SMEs: Bipasha Banerjee, Sara Ahmadi

Team:

Kanad Naleshwarkar, Gayatri Bhatambarekar, Zeel Desai, Aishwarya Kumaran, Adithya Harish Srinivasan Manikandan, Shadab Haque

VIRGINIA TECH

5 December 2023, Blacksburg, VA, 24061, USA

Classification - Summary of tasks

- Evaluation of previous implementation.
- Fine-tuned the SciBERT[1] model on the 50k abstract-label dataset.
- Addressed label imbalance problem by assigning weightage to each label.
- Classifier model now provides output in accordance with ProQuest level-2 subject categories.
- Implemented batch processing to speed up the inference from new model.

SciBERT Classification - Results

- Average F1 score on test set using SciBERT[1]: 0.74
- Average loss (BCEWithLogitsLoss[2]): 0.72

Classification - Sample Output

Sample ETD ID: **75199**

Title: *"Faith and politics: The socio-political discourses engaged by Mexican ex-voto paintings from the nineteenth-century and beyond."*

Abstract: *"The Universalis Ecclesiae of 1508 authorized Spanish colonization of the Americas in return for the conversion of native populations to Christianity. From its inception therefore, the Mexican nation lived an alliance between Church and State. This alliance promoted the transfer of Castilian Catholicism to American shores. Catholic practices, specifically the ex-voto tradition, visualize this intermingling of religion and politics. The ex-voto is a devotional painting that expresses…"*

Model output: *{'History': '0.63', 'Fine and performing arts': '0.44'}*

SciBERT Classification Statistics

- Time taken to fetch 1000 title-abstract pairs from the APIs: 73.7 sec
- Time taken for inference on 1000 title-abstract pairs: 2.4 sec

Classification using LLaMA 2

- Approach to overcome class imbalance:
  - Oversampling
- Removed all the rows having no abstracts.
- The structure of the prompts is as follows:

```
'Below is an abstract of an article. Categorize it appropriately.\n\nCategories:\n\nENGINEERING\nHEALTH AND MEDICAL SCIENCES
| EDUCATION\nPHILOSOPHY AND RELIGION\nBUSINESS\nCOMMUNICATIONS AND INFORMATION SCIENCES\nHISTORY\nFINE AND PERFORMING ARTS\n
COMMUNICATIONS AND INFORMATION SCIENCES | INTERDISCIPLINARY\nLANGUAGE AND LITERATURE\nEDUCATION\nBEHAVIORAL SCIENCES\nAGRICU
LTURE\nSOCIAL SCIENCES\nBIOLOGICAL SCIENCES\nARCHITECTURE\nGEOSCIENCES\nMATHEMATICAL AND PHYSICAL SCIENCES\nENGINEERING | AR
CHITECTURE | ENVIRONMENTAL SCIENCES\nBEHAVIORAL SCIENCES | EDUCATION\n\nAbstract:\nvast individual differences exist at both
the personality level and the brain level that can influence anxiety as a general tendency...\n\n### Category:\nBEHAVIORAL S
CIENCES\n\n### End'
```

Classification using LLaMA 2

- Fine-tuned Llama-2-7b-chat-hf.
- Trained it on 150000 oversampled data for a total of 1 epoch.

Metric	**Value (%)**
Accuracy	48.31
Precision	68.74
Recall	48.31
F1 Score	52.85

Table 1: LLaMA 2 classification evaluation metrics.

LLaMA 2 Classification Sample Output

Article:
abstract
self-compassion and self-forgiveness appear to have much in common, as both relate to one's self-regard during challenging c
ircumstances; however, their empirical relationship is largely yet to be explored. the present study examines theoretical an
d empirical areas of overlap and proposes a theory of their relationship, including its possible impact on health. self-comp
assion and self-forgiveness were proposed to have a direct relationship that may be mediated by reduction in rumination, sha
me, and experiential avoidance. these factors together were also hypothesized to have a positive impact on health functionin
g. the current study tested these models in a sample of undergraduate students (n = 199). in parallel mediation analysis, se
lfcompassion and self-forgiveness were related to one another and this association was partially mediated by shame, only. as
such, neither rumination nor experiential avoidance were included in subsequent analyses. in serial mediation analyses, self
-compassion, shame, and selfforgiveness were found to affect health outcomes in various ways. for psychological distress, th
e association of self-compassion was partially mediated by shame and self-forgiveness, in an indistinguishable fashion. for
mental health status, self-compassion was found to have a significant direct effect, only. for both somatic symptoms (full m
ediation) and physical health status (indirect only effect), the association of self-compassion operated through self-forgiv
eness both alone and linked with shame. as these two emerging areas in the psychological literature continue to mature, rese
archers should prioritize both integration between and nuances within these constructs in order to develop a more complete u
nderstanding of self-compassion, selfforgiveness, and their implications for health.

Generated Category:
BEHAVIORAL SCIENCES

End
Ground Truth Summary:
BEHAVIORAL SCIENCES

LLaMA 2 Classification Sample Output

Article:
in this paper, i argue that the marginal hybrids in the macclesfield psalter functioned as loci of ethical cultivation. the marginal hybrids, i contend, functioned on three levels. on the practical level, they served as aides-mémoire, playing a critical role in the overall visual system of the manuscript and its design in fulfillment of the basic principles of medieval mnemotechnique. on the spiritual level, they served as sites for and aids to meditatio, the second step in the process of lectio divina. as facilitators of meditatio, the marginal hybrids assisted the reader in his quest to discover the deeper significance of the scriptures and, by so doing, to strive for what french scholar jacques hourlier called connaturalness with god. because of their roles in both memorization and meditation, the marginal hybrids also operated on an ethical level as loci and machinae of moral cultivation. they were the means by which the reader domesticated the text in his mind and heart and confronted his inner self in order to reshape himself in the image of god.

Generated Category:
LANGUAGE AND LITERATURE

End
Ground Truth Summary:
FINE AND PERFORMING ARTS

Classification - Evaluation

- SciBERT model for classification outperformed LLaMA 2 (F1 score **0.74** for SciBERT vs. **0.52** for LLaMA 2).
- SciBERT was significantly faster to get inference on title-abstract pairs than LLaMA 2.

Classification - Future Work

- Speeding up inference process using Kafka.
- Running classification on chapter summaries.

Summarization - Summary of tasks

We have used 5 models to evaluate generated summaries:

1. TextRank (Non-transformer model)
2. LexRank (Non-transformer model)
3. LSA (Non-transformer model)
4. BigbirdPegasus (Transformer model)
5. LLaMA 2 7B

We have used multiple ETDs as input and we are able to successfully generate summaries of each ETD.

Summarization - summary of tasks

- Analysed the previous year team's code and improved the models.
- Worked on 5 models to generate summaries from input PDFs.
- Made summaries cleaner by removing unnecessary words (stopwords) for all models.
- Checked all the model performance using ROUGE score for precision, recall, and F1 metrics[4].

Instruction Tuning for LLaMA 2

- The arXiv and Booksum datasets were used for fine-tuning.
- The model used was LLaMA-2-7B.
- The maximum sequence used was 4096.
- For removing noise from the arXiv dataset:
  - Removed citations @xcite or @xmath.
  - Removed LaTeX notations.
  - Removed references starting with arXiv or DOI.
- The text field contains a prompt, article and the abstract.
- Text field is given as input to the model for fine-tuning.

Summarization - TextRank

TextRank[3] is an algorithm used in Natural Language Processing (NLP) for extractive summarization and keyword extraction. It's based on the PageRank algorithm, which was originally used by Google to rank web pages in search results. We used the chapter titled “Discipline-Independent Text Information Extraction from Heterogeneous Styled References Using Knowledge from the Web”. The summarized output is shown in Fig. 1.


```
{"summarisation": "Chapter 1 Introduction 1.1 1.1.1 scholarly digital libraries, citation/reference analysis support patrons leads suitable
services. Citation relations digital objects collections identified. Examples citation analysis include: metadata extraction
[2,10,12,14,32,35,56,96], citation matching [47, 69, 94], citation/co-authorship network construction [3, 5], citation/coauthorship network
analysis [18,88]. Due success machine learning, natural language processing, web mining, researchers envision Machine Reading [23, 76], goes
beyond domain-independent unsupervised text understanding. aim help move toward vision Machine Reading using knowledge derived web mining,
along machine learning techniques support vector machines (SVMs) conditional random fields (CRFs). Further, learn analysis collections
references appear large documents, opposed working individual references. Sung H. Park give focus, time ensuring generality findings, study
references two representative digital library environments. cases, one finds scholarly digital documents including heterogeneously styled
references contributed many disciplines. particular, study ETD-db digital library system1 arXiv system2 . ETD-db digital library system
supports higher education, adopted/adapted hundreds locations. supports submitting, reviewing, cataloging, managing Electronic Theses
Dissertations (ETDs) local collections, many connected Networked Digital Library Theses Dissertations (NDLTD, see www.ndltd.org). hand, arXiv
system centralized open electronic-publication system, run Cornell University Library, pre-prints reprints Physics, Mathematics, Computer
Science, Quantitative Biology, Quantitative Finance, Statistics. working digital libraries aim aggregate publications across groups
disciplines [79]. instance, consider Virginia Tech\u2019s ETD-db digital library, instantiation ETD-db system originally developed Virginia
Tech. order citation analysis algorithm scalable, able extract information heterogeneous collections, putting results suitable canonical
forms. many attempts segment parts digital content digital libraries (e.g., tables [55, 74], figures [57], chemical formula [86],
acknowledgments [41], etc.) Entlich et al. (1997) [22], CORE project, aimed serve needs chemists 1 http://scholar.lib.vt.edu/ETD-db/
http://www.arxiv.org 3", "algorithm_used": "textrank"}
```


Fig. 1: Summary using TextRank

Summarization - LexRank

LexRank is a graph-based algorithm used in text summarization to identify and extract key sentences from a document. It's an extractive summarization technique, meaning it selects important sentences from the input text to create a summary. The summarized output is shown in Fig. 2.


```
{"summarisation": "Chapter 1 Introduction 1.1 1.1.1 scholarly digital libraries, citation/reference analysis support patrons leads suitable
services. Examples citation analysis include: metadata extraction [2,10,12,14,32,35,56,96], citation matching [47, 69, 94], citation/co-
authorship network construction [3, 5], citation/coauthorship network analysis [18,88]. studies important, since scholars use citations
support discussion argumentation, since collections references aid identifying relationships among works. Text Processing, e.g., Support
Machine Reading Due success machine learning, natural language processing, web mining, researchers envision Machine Reading [23, 76], goes
beyond domain-independent unsupervised text understanding. Further, learn analysis collections references appear large documents, opposed
working individual references. However, requires additional processing, e.g., chapter section references, bibliographies, footnotes. Sung H.
Park requirements related research broad. cases, one finds scholarly digital documents including heterogeneously styled references contributed
many disciplines. particular, study ETD-db digital library system1 arXiv system2 . ETD-db digital library system supports higher education,
adopted/adapted hundreds locations. supports submitting, reviewing, cataloging, managing Electronic Theses Dissertations (ETDs) local
collections, many connected Networked Digital Library Theses Dissertations (NDLTD, see www.ndltd.org). Scalability across Disciplines working
digital libraries aim aggregate publications across groups disciplines [79]. Extracting references collections difficult. order citation
analysis algorithm scalable, able extract information heterogeneous collections, putting results suitable canonical forms. operate across
disciplines reference styles, handle broad range reference surface forms [84]. http://scholar.lib.vt.edu/ETD-db/ http://www.arxiv.org 3",
"algorithm_used": "lexrank"}
```


Fig. 2: Summary using LexRank

Summarization - LSA

Latent Semantic Analysis (LSA) can be used for text summarization, particularly for extractive summarization, where the goal is to select a subset of sentences from a document to create a summary. The summarized output is shown in Fig. 3.


```
{"summarisation": "scholarly digital libraries, citation/reference analysis support patrons leads suitable services. Citation relations
digital objects collections identified. Examples citation analysis include: metadata extraction [2,10,12,14,32,35,56,96], citation matching
[47, 69, 94], citation/co-authorship network construction [3, 5], citation/coauthorship network analysis [18,88]. studies important, since
scholars use citations support discussion argumentation, since collections references aid identifying relationships among works. Text
Processing, e.g., Support Machine Reading Due success machine learning, natural language processing, web mining, researchers envision Machine
Reading [23, 76], goes beyond domain-independent unsupervised text understanding. aim help move toward vision Machine Reading using knowledge
derived web mining, along machine learning techniques support vector machines (SVMs) conditional random fields (CRFs). overcome limitations
current citation information extraction techniques, work domain-specific supervised manner. However, requires additional processing, e.g.,
chapter section references, bibliographies, footnotes. give focus, time ensuring generality findings, study references two representative
digital library environments. cases, one finds scholarly digital documents including heterogeneously styled references contributed many
disciplines. ETD-db digital library system supports higher education, adopted/adapted hundreds locations. supports submitting, reviewing,
cataloging, managing Electronic Theses Dissertations (ETDs) local collections, many connected Networked Digital Library Theses Dissertations
(NDLTD, see www.ndltd.org). hand, arXiv system centralized open electronic-publication system, run Cornell University Library, pre-prints
reprints Physics, Mathematics, Computer Science, Quantitative Biology, Quantitative Finance, Statistics. working digital libraries aim
aggregate publications across groups disciplines [79]. instance, consider Virginia Tech\u2019s ETD-db digital library, instantiation ETD-db
system originally developed Virginia Tech. Another example arXiv, 662,023 e-prints 7 large disciplines, covering 148 sub-categories4 March
2011. number reference extraction analysis approaches systems make use domain knowledge, specialized area, computing chemistry. order citation
analysis algorithm scalable, able extract information heterogeneous collections, putting results suitable canonical forms. Entlich et al.
(1997) [22], CORE project, aimed serve needs chemists 1", "algorithm_used": "lsa"}
```


Fig. 3: Summary using LSA

Summarization - BigBirdPegasus

BigBird[5] is designed to handle long sequences, making it well-suited for tasks that involve lengthy documents and text summarization. It is particularly useful for extractive summarization tasks where the goal is to select important sentences or passages from a document to create a summary.The summarized output is shown in Fig. 4.

{"summarisation": "this paper presents an analysis of the relationship between references and digital libraries .<n> we show that it is possible to identify links between large collections of reference forms from heterogeneously appearing digital documents .<n> furthermore , our study shows that citation analysis can be a powerful way to support higher education through identification of appropriate relations between scholarly objects . <n> [[section]] in this chapter , we give an overview of our research on how to support higher education through identification of appropriate relationships among large collections of reference forms .<n> first , we discuss some aspects of our research for which we have presented results (chapters 1 - 7) .<n> second , we focus on two main issues : firstly , what do we mean by large collections of reference forms ? secondly , are there any potential ways to support higher education through identification of appropriate relations between small collections of reference forms ?", "algorithm_used": "bigbird-pegasus"}

Fig. 4: Summary using BigbirdPegasus

Instruction Tuning for Summarization

Ground Truth Summary:
the availability of new data sources on human mobility is opening new avenues for investigating the interplay of social networks human mobility and dynamical processes such as epidemic spreading here we analyze data on the time resolved face to face proximity of individuals in large scale real world scenarios we compare two settings with very different properties scientific conference and long running museum exhibition we track the behavioral networks of face to face proximity and characterize them from both static and dynamic point of view exposing differences and similarities we use our data to investigate the dynamics of susceptible infected model for epidemic spreading that unfolds on the dynamical networks of human proximity the spreading patterns are markedly different for the conference and the museum case and they are strongly impacted by the causal structure of the network data deeper study of the spreading paths shows that the mere knowledge of static aggregated networks would lead to erroneous conclusions about the transmission paths on the dynamical networks

Generated Summary:
the analysis of time resolved networks of interactions among individuals provides new insights on the nature of the social interactions and on the structural properties of the networks we have collected rich datasets of face to face interactions between individuals using novel radio frequency identification technology we have shown how the topology of the aggregated networks of interactions differs in two very different settings conference and museum the network dynamics plays important role in the spreading of simple dynamical processes such as the transmission of piece of information or of an infectious agent the availability of time resolved data reveals interesting differences between the two settings and unveils the importance of the role played by causality in the spreading process we have also shown that although time resolved networks are complex and dynamic their structures are often similar to those of static networks we investigate these issues by comparing the time resolved data to static ones collected from literature and by analyzing the role played by causality in the unfolding of dynamical processes

Evaluation Metrics - ROUGE-1 score

We used the ROUGE score method to assess the quality of the summaries generated by our models. The results of this evaluation are presented in Table 2.

Model Name	F1	Recall	Precision
Text Rank	0.28	0.31	0.25
Lex Rank	0.26	0.34	0.22
LSA	0.29	0.38	0.23
Big Bird Pegasus	0.235	0.206	0.73

Table 2: Evaluating models using ROUGE-1 scores

Evaluation Metrics - ROUGE-2 score

Table 3 depicts ROUGE-2 scores. The reference summary is manually generated summary.

Model Name	F1	Recall	Precision
Text Rank	0.092	0.103	0.083
Lex Rank	0.081	0.103	0.067
LSA	0.08	0.10	0.60
Big Bird Pegasus	0.029	0.25	0.34

Table 3: Evaluating models using ROUGE-2 scores

Evaluation Metrics - ROUGE-L score

Table 4 depicts ROUGE-L scores. The reference summary is manually generated summary.

Model Name	F1	Recall	Precision
Text Rank	0.156	0.175	0.141
Lex Rank	0.141	0.180	0.117
LSA	0.14	0.18	0.11
Big Bird Pegasus	0.141	0.123	0.164

Table 4: Evaluating models using ROUGE-L scores

Evaluation - ROUGE-1 score (abstract as reference summary)

Table 5 depicts ROUGE-1 scores. The reference summary is abstract of the chapter.

Model Name	F1	Recall	Precision
Text Rank	0.56	0.45	0.76
Lex Rank	0.51	0.39	0.74
LSA	0.46	0.56	0.70
Big Bird Pegasus	0.10	0.08	0.55

Table 5: Evaluating models with respect to abstract using ROUGE-1 scores

Instruction-tuning for Summarization

- ROUGE scores for the previous summary:

```
ROUGE1:
  Precision: 0.5260
  Recall: 0.5449
  F-measure: 0.5353

ROUGE2:
  Precision: 0.1570
  Recall: 0.1627
  F-measure: 0.1598

ROUGEL:
  Precision: 0.2601
  Recall: 0.2695
  F-measure: 0.2647
```

- Train with more maximum sequence tokens for improvement in evaluation score.

Evaluation Results

- We used ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metrics to check how good our text summaries are. These metrics helped us to see if the summaries are good or not based on ROUGE-1, ROUGE-2 and ROUGE-L scores indicating precision, recall and F1 score. TextRank model outperformed other methods and gave the best results.
- When we compared generated summaries to manually generated summaries we found that BigBirdPegasus model outperformed all the other models followed by the TextRank.
- When we compared generated summaries with the abstract of the chapter we found that TextRank performed a lot better than other models.
- Apart from using ROUGE, we manually reviewed the summaries, and summaries generated by TextRank were more appealing. Its summaries looked better to us, and it was also the quickest to generate summaries.
- Currently we have decided to use TextRank as our main model for summarizing ETDs.

Summarization - Future Work

1. Model Improvements:

BigBirdPegasus: We did preprocessing, tokenization to remove stop words, but one could improve the accuracy further to generate more comprehensive and informative summaries.

LongFormer Model: It could be utilized and better fine-tuning can be done in order to increase the accuracy of the summaries. It could generate summaries that are not only informative but also meaningful and fluent.

LLaMA 2 Model: There could be more experimentation done with the structure of the input text while fine-tuning. It can also be run for a larger number of epochs.

2. Furthermore, to improve the model performance, the dataset size can be increased and extensive model tuning can be performed.

3. There can be an improvement to refine and develop evaluation metrics that provide a comprehensive assessment of summarization quality. Metrics like coherence, informativeness, and coverage could be utilized beyond traditional ROUGE scores.

References

[1] I. Beltagy, K. Lo, and A. Cohan. SciBERT: A Pretrained Language Model for Scientific Text. arXiv preprint arXiv:1903.10676, 2019. URL https://ar5iv.org/abs/1903.10676. (Accessed Dec 13, 2023).

[2] S. Imambi, K. B. Prakash, and G. R. Kanagachidambaresan. Pytorch. In K. B. Prakash and G. R. Kanagachidambaresan, editors, Programming with TensorFlow: Solution for Edge Computing Applications, pages 87–104. Springer International Publishing, Cham, 2021. URL https://doi.org/10.1007/978-3-030-57077-4_10. (Accessed Dec 13, 2023)

[3] R. Mihalcea and P. Tarau. TextRank: Bringing Order into Text. In D. Lin and D. Wu, editors, Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pages 404–411, Barcelona, Spain, 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-3252. (Accessed Dec 17, 2023).

[4] R. Nallapati, B. Zhou, C. Gulcehre, and B. Xiang. Abstractive text summarization using sequence-to-sequence RNNs and beyond. arXiv:1602.06023, 2016. URL https://arxiv.org/abs/1602.06023. (Accessed Dec 13, 2023).

[5] M. Zaheer, G. Guruganesh, K. A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang, and A. Ahmed. Big Bird: Transformers for Longer Sequences. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, Advances in Neural Information Processing Systems, volume 33, pages 17283–17297. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/c8512d142a2d849725f31a9a7a361ab9-Paper.pdf. (Accessed Dec 17, 2023).

THANK YOU