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

5 December 2023, Blacksburg, VA, 24061, USA
Classification - Summary of tasks

- Evaluation of previous implementation.
- Fine-tuned the SciBERT\cite{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\textsuperscript{[1]}: 0.74
- Average loss (BCEWithLogitsLoss\textsuperscript{[2]}): 0.72
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.

Categories: ENGINEERING | HEALTH AND MEDICAL SCIENCES | EDUCATION | PHILOSOPHY AND RELIGION | BUSINESS | COMMUNICATIONS AND INFORMATION SCIENCES | HISTORY | FINE AND PERFORMING ARTS | COMMUNICATIONS AND INFORMATION SCIENCES | INTERDISCIPLINARY | LANGUAGE AND LITERATURE | EDUCATION | BEHAVIORAL SCIENCES | AGRICULTURE | SOCIAL SCIENCES | BIOLOGICAL SCIENCES | ARCHITECTURE | GEOSCIENCES | MATHEMATICAL AND PHYSICAL SCIENCES | ENGINEERING | ARCHITECTURE | ENVIRONMENTAL SCIENCES | BEHAVIORAL SCIENCES | EDUCATION

Abstract: A vast individual differences exist at both the personality level and the brain level that can influence anxiety as a general tendency..."
Classification using LLaMA 2

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

<table>
<thead>
<tr>
<th>Metric</th>
<th>Value (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>48.31</td>
</tr>
<tr>
<td>Precision</td>
<td>68.74</td>
</tr>
<tr>
<td>Recall</td>
<td>48.31</td>
</tr>
<tr>
<td>F1 Score</td>
<td>52.85</td>
</tr>
</tbody>
</table>

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 circumstances; however, their empirical relationship is largely yet to be explored. The present study examines theoretical and empirical areas of overlap and proposes a theory of their relationship, including its possible impact on health. Self-compassion and self-forgiveness were proposed to have a direct relationship that may be mediated by reduction in rumination, shame, and experiential avoidance. These factors together were also hypothesized to have a positive impact on health functioning. The current study tested these models in a sample of undergraduate students (n = 199). In parallel mediation analysis, self-compassion 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 self-forgiveness were found to affect health outcomes in various ways. For psychological distress, the 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 mediation) and physical health status (indirect only effect), the association of self-compassion operated through self-forgiveness both alone and linked with shame. As these two emerging areas in the psychological literature continue to mature, researchers should prioritize both integration between and nuances within these constructs in order to develop a more complete understanding of self-compassion, self-forgiveness, 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 connaturaleness 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.
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.
TextRank 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.

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.

Fig. 2: Summary using LexRank
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.

Fig. 3: Summary using LSA
BigBird\cite{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.

```json
{"summarisation": "this paper presents an analysis of the relationship between references and digital libraries. We show that it is possible to identify links between large collections of reference forms from heterogeneously appearing digital documents. Furthermore, our study shows that citation analysis can be a powerful way to support higher education through identification of appropriate relationships between scholarly objects. 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. First, we discuss some aspects of our research for which we have presented results (chapters 1-7). 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 scientifi c conference and long running museum exhibition we track the behavioral networks of face to face proximity and characterize t hem 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 spread ing patterns are markedly different for the conference and the museum case and they are strongly impacted by the causal struc ture of the network data deeper study of the spreading paths shows that the mere knowledge of static aggregated networks woul d 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 in teractions and on the structural properties of the networks we have collected rich datasets of face to face interactions betw een individuals using novel radio frequency identification technology we have shown how the topology of the aggregated networ ks 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 av ailability of time resolved data reveals interesting differences between the two settings and unveils the importance of the r ole played by causality in the spreading process we have also shown that although time resolved networks are complex and dyna mic 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 pro cesses
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.

<table>
<thead>
<tr>
<th>Model Name</th>
<th>F1</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text Rank</td>
<td>0.28</td>
<td>0.31</td>
<td>0.25</td>
</tr>
<tr>
<td>Lex Rank</td>
<td>0.26</td>
<td>0.34</td>
<td>0.22</td>
</tr>
<tr>
<td>LSA</td>
<td>0.29</td>
<td>0.38</td>
<td>0.23</td>
</tr>
<tr>
<td>Big Bird Pegasus</td>
<td>0.235</td>
<td>0.206</td>
<td>0.73</td>
</tr>
</tbody>
</table>

Table 2: Evaluating models using ROUGE-1 scores
Table 3 depicts ROUGE-2 scores. The reference summary is manually generated summary.

<table>
<thead>
<tr>
<th>Model Name</th>
<th>F1</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text Rank</td>
<td>0.092</td>
<td>0.103</td>
<td>0.083</td>
</tr>
<tr>
<td>Lex Rank</td>
<td>0.081</td>
<td>0.103</td>
<td>0.067</td>
</tr>
<tr>
<td>LSA</td>
<td>0.08</td>
<td>0.10</td>
<td>0.60</td>
</tr>
<tr>
<td>Big Bird Pegasus</td>
<td>0.029</td>
<td>0.25</td>
<td>0.34</td>
</tr>
</tbody>
</table>

Table 3: Evaluating models using ROUGE-2 scores
Table 4 depicts ROUGE-L scores. The reference summary is manually generated summary.

<table>
<thead>
<tr>
<th>Model Name</th>
<th>F1</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text Rank</td>
<td>0.156</td>
<td>0.175</td>
<td>0.141</td>
</tr>
<tr>
<td>Lex Rank</td>
<td>0.141</td>
<td>0.180</td>
<td>0.117</td>
</tr>
<tr>
<td>LSA</td>
<td>0.14</td>
<td>0.18</td>
<td>0.11</td>
</tr>
<tr>
<td>Big Bird Pegasus</td>
<td>0.141</td>
<td>0.123</td>
<td>0.164</td>
</tr>
</tbody>
</table>
Evaluation - ROUGE-1 score (abstract as reference summary)

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

<table>
<thead>
<tr>
<th>Model Name</th>
<th>F1</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text Rank</td>
<td>0.56</td>
<td>0.45</td>
<td>0.76</td>
</tr>
<tr>
<td>Lex Rank</td>
<td>0.51</td>
<td>0.39</td>
<td>0.74</td>
</tr>
<tr>
<td>LSA</td>
<td>0.46</td>
<td>0.56</td>
<td>0.70</td>
</tr>
<tr>
<td>Big Bird Pegasus</td>
<td>0.10</td>
<td>0.08</td>
<td>0.55</td>
</tr>
</tbody>
</table>

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.
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


THANK YOU