Automated Hypotheses Generation in COVID-19 Studies

Meghana Maringanti – University of Arizona

Abstract

This paper explores automated hypotheses generation using a dataset of research papers on COVID-19. In this project we utilized cosine similarity to group research titles and abstracts, creating meaningful contexts for hypothesis generation. A language model was employed to generate and refine ideas, which were then evaluated for relevance and feasibility. This approach aims to streamline the hypothesis generation process, aiding researchers in identifying innovative and feasible research directions.

1 Introduction

Due to the COVID-19 pandemic,, it came to light how important it is to come up with new research ideas quickly. With so many studies being published, it became hard for researchers to find useful information and figure out what areas still needed attention. Traditional ways of generating research ideas take time and effort, which isn't ideal when the world needs solutions fast. This makes automation a valuable tool to help speed up the process and support scientists in tackling big challenges.

In this project, we explored NLP techniques to automatically generate research ideas from a given dataset. We used cosine similarity to group similar titles and abstracts into clusters, which made it easier to find patterns and connections. These clusters provided a starting point for generating hypotheses using a language model. The model created ideas and improved them through refinement, ensuring they were relevant and practical.

This process can help researchers discover new directions in their work. While this study focused on COVID-19, the same methods can be applied to other areas. This paper explains the process, the results I achieved, and how this approach could improve the way research is done in the future.

2 Methodology

2.1 Dataset

The dataset used in this study is the COVID-19 and SARS-CoV-2 Research Papers Dataset (Salieh, 2023), and was obtained from Hugging Face.

This dataset contains 1,035 rows and 6 columns and provides information on paper titles, abstracts, journals, publication dates, authors, and DOIs.

2.2 Pre-processing

To ensure that the data is ready for analysis several pre-processing steps were applied. For this analysis, we mainly worked with the title and abstract of the papers. First, missing values were handled and they were then cleaned by converting all text to lowercase, removing non-alphabetic characters, and tokenizing the text into individual words. Common stop-words (e.g., "the", "and", "is") were also removed to reduce noise and focus on meaningful words.

After pre-processing the average length of abstracts we used is around 140.

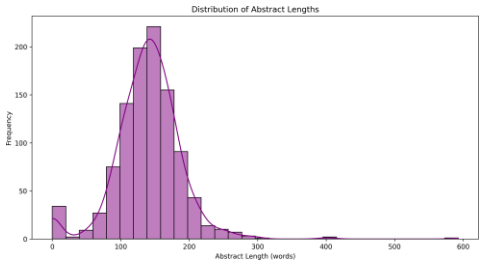

Figure 1: This image shows the distribution of abstract lengths after pre-processing.

Then we applied cosine similarity to this processed title and abstract and grouped into clusters. These clusters were then provided as context to generate new hypotheses.

2.3 Model

Meta-Llama-3.1-8B-Instruct-Turbo:

This model is used for both idea generation and refinement. It is developed by Meta AI and is designed to perform a variety of natural language processing tasks with high efficiency, including text generation, summarization, and refinement, by following the specific instructions provided in the prompts.

2.4 Idea Generation

We accessed the model mentioned above through **Together AI** API and utilized it to analyze the clusters created above and generate novel research ideas. The input corpus consisted of the combined *title* and *abstract* of each research paper in a similar cluster, providing context for generating hypotheses related to COVID-19 research.

The generation prompt was designed to extract valuable insights by focusing on three key aspects:

1. **Gaps in Existing Research:**

   Identifying areas where the current research has not sufficiently addressed specific questions or problems.

2. **Potential Unexplored Areas:**

   Suggesting new areas that have yet to be fully explored and might offer significant contributions to the field.

3. **Improvements to Existing Ideas:**

   Proposing enhancements to current methodologies or theories that could lead to better outcomes.

2.5 Research Question Refinement

After the initial idea generation, we refined the generated ideas to ensure they were actionable and scientifically valid. The same model above is used for this step, using a more structured refinement prompt to improve the clarity, specificity, and feasibility of each idea.

The refinement prompt required the model to:

- **Refine each generated idea:**

  To include critical elements such as a clear research question and hypotheses.

- **Categorize into themes:**

  Gaps in Existing Research, Potential Unexplored Areas, or Improvements to Existing Ideas.

- **Concise and structured responses**:

  To ensure that the ideas were actionable and scientifically sound.

The refined ideas were presented in a clear format, making them suitable for further development and exploration within the context of ongoing research in the field.

All these ideas are stored in a text file and then processed to remove special symbols and numerical and then stored into a .csv file with columns *category* and *Research questions*

2.6 Evaluation

Two models which were different than the model used for the generation and the refinement of the research questions were utilized to assess the relevance and feasibility of the generated research ideas

2.6.1 Model

1. **SentenceTransformer model:**

   This was used for calculating the relevance scores. Using this model I generated embeddings for both the research corpus (cleaned title and abstract) and the generated research questions.

   The cosine similarity between the embeddings of the research questions and the corpus was then calculated.

   The relevance score for each idea was determined by selecting the maximum cosine similarity value between the research question's embedding and all dataset embeddings.

   This method effectively captures how closely each idea aligns with the content of the research corpus.

2. **T5-small model:**

   It is accessed through the Hugging Face 'pipeline' for text classification and, was used to evaluate the feasibility of the generated research ideas.

The feasibility score for each idea was obtained by prompting the model to assess the feasibility of the research question. The model outputs a score indicating the likelihood of the idea being feasible.

2.6.2 Analysis

The relevance and feasibility scores were visualized using histograms to better understand their distribution. The histograms display the spread of both relevance and feasibility scores, with Kernel Density Estimation (KDE) overlaid to highlight the shape of the distributions. This visualization helps in understanding the general trend and variance of the scores for the generated research ideas.

The Visualization shows

a. Relevance Scores:

The distribution of how closely the generated research questions align with the context of the input research corpus.

b. Feasibility Scores:

The distribution shows the likelihood of the research ideas being feasible, with higher scores indicating more feasible ideas.

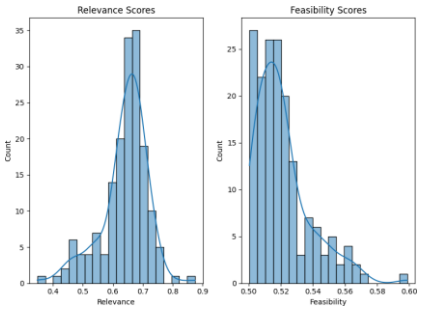

Figure 2: This image shows the distribution of abstract lengths after pre-processing.

Below are the observations based on the scores:

1. **Relevance Scores:**

   The distribution of relevance scores appears to be approximately normal, with a central peak around 0.65.

   The majority of the relevance scores are concentrated in the range of 0.60–0.70, indicating that most of the hypotheses are evaluated as moderately relevant.

   There are relatively fewer scores below 0.50 and above 0.75, suggesting that extreme relevance ratings (either very low or very high) are uncommon.

2. **Feasibility Scores:**

   The distribution of feasibility scores is skewed to the right, with a peak near 0.52 and a tail extending toward higher values.

   The majority of feasibility scores fall within the 0.50–0.55 range, showing a tendency toward moderate feasibility assessments.

   Very few scores are observed above 0.58, indicating that highly feasible hypotheses are relatively rare.

   The right-skewed nature suggests that while many hypotheses are moderately feasible, there are fewer high-scoring outliers.

Limitations

1. **Fixed and Focused Dataset:**

   The project uses a static dataset and is limited to COVID-19 topics. Hence it reduces generalization for other research areas and also lacks real-time updates or broader applicability.

2. **No Standard Evaluation Metrics:**

   Hypothesis quality lacks standardized metrics, making quantitative assessment difficult.

3. **Duplication Challenges:**

   Cosine similarity may miss nuanced overlaps, leading to redundant hypotheses.

4. **Lack of Validation:**

   Generated ideas require domain experts for refinement and practical validation and the Feasibility scores we included rely on LLMs output, which reduces accuracy..

These limitations highlight the need for real-time data, expert reviews, and robust evaluation methods.

Ethics Statement

This project adheres to ethical principles by ensuring that all data used is sourced from publicly available and reputable repositories, respecting their terms of use. No personally identifiable information (PII) or sensitive data is involved.

The project aims to support scientific research by identifying gaps and generating hypotheses without altering or misrepresenting the original research content.

Any automated processes comply with the guidelines of the platforms accessed, and proper citations are included to credit original authors.

The generated hypotheses are suggestions and should be further validated by domain experts.

Conclusion

This project explores an approach to automatically generating and evaluating hypotheses in three main categories, Gaps in Existing Research, Potential Unexplored Areas, and Improvements to Existing research in the context of COVID-19 research and then evaluate them using large language models.

The analysis of relevance and feasibility scores reveals that most hypotheses fall within moderate ranges, with relevance scores clustering around 0.65 and feasibility scores peaking near 0.52. This indicates that while the generated hypotheses are generally relevant and feasible, exceptionally high-quality hypotheses are less common.

The methodology used in this project is scalable and can be used in identifying research gaps and generating new ideas in any domains. However, the study also shows the challenges faced if we rely solely on automated systems for hypothesis evaluation.

We need to enhance accuracy through fine-tuning the models, provide a real-time dataset integration rather than a static dataset, and also need collaboration with subject-matter experts to refine and validate results.

Future Scope

The future scope of this project involves improving the model to automatically fetch all related research papers directly from online repositories rather than relying on a static dataset. By integrating APIs or web scraping tools for platforms like PubMed, arXiv, or Semantic Scholar, the system can stay updated with the latest research. This approach will enable real-time analysis of newly published research, and improve the identification of gaps and trends in the literature.

References

[1] F. G. Salieh, COVID-19 and SARS-CoV-2 Research Papers Dataset, Version 1.0, Hugging Face, 2023. [Online]. Available: https://huggingface.co/datasets/falah/research_paper_in_ml