Can AI Assist in Paper Writing? Automatic Keyword Extraction, Hypothesis Generation, and Topic Modeling: Evaluating the Feasibility

Rony Macwan
College of Information Science
University of Arizona
ronymacwan@arizona.edu

Abstract

The use of artificial intelligence (AI) for Scientific Discovery has gained significant attention in recent years. This project explores the feasibility of using Natural Language Processing (NLP) techniques to facilitate automatic Scientific Discovery, with a specific focus on keyword generation, hypothesis generation, and topic modeling for research topics. By using NLP models like BERT and Together AI, the project evaluates the effectiveness of these tools in automating tasks traditionally handled by researchers. The goal is to evaluate how these tools can streamline research workflows, accelerate the discovery process, and inspire innovative approaches to scientific exploration. The findings of this project will offer valuable insight into how AI can help in scientific communication, such as research papers and journal articles, through Automatic Scientific Discovery. Additionally, the study will critically examine the reliability and limitations of these AI tools.

1 Introduction

Scientific Discovery plays a crucial role in advancing knowledge and solving important problems. Traditionally, Scientific Discovery has relied on the ingenuity and creativity of researchers to identify key problems, generate hypotheses, and craft impactful studies. However, with the exponential growth of information in the digital age, it has become increasingly challenging for researchers to navigate vast amounts of literature, identify emerging trends, and uncover novel ideas. This has paved the way for the integration of Artificial Intelligence (AI) as a tool to assist in the Scientific Discovery process. AI-powered systems, particularly those that use natural language processing (NLP), have gained traction due to their ability to process and analyze large datasets with remarkable speed and accuracy. Tools such as ChatGPT, LLaMA, and Retrieval-Augmented Generation (RAG) systems have been used to assist researchers in various capacities. These tools can generate coherent text, summarize papers, suggest relevant topics, and even propose new ideas based on existing literature. Their accessibility and efficiency have made them attractive options for improving productivity in academic and research settings. Although these advances are promising, it is crucial to critically evaluate their reliability and limitations. Relying on AI-generated output for tasks such as keyword generation or ideation introduces concerns about accuracy, originality, and ethical considerations. The ability of AI to create meaningful and novel contributions to scientific discourse must be analyzed to determine whether these tools can genuinely enhance the discovery process or if they risk introducing noise into research. A structured evaluation is needed to assess the effectiveness of these systems and ensure they support, rather than reduce, the quality of scientific research.

This project aims to explore the role of NLP techniques in supporting Automatic Scientific Discovery, with a focus on keyword generation, hypothesis generation, and topic modeling. By analyzing the capabilities and limitations of these tools, the study seeks to provide insights into their potential to change how researchers approach scientific exploration.

2 Methods

The first task was to automatically generate keywords for research papers. I began by collecting a dataset of 200 research papers focused on Machine Learning from Hugging Face. These papers were then preprocessed to create summaries, which formed the basis for our keyword extraction. To achieve this, I implemented two distinct approaches: BERT embeddings combined with TF-IDF and keyword generation using the Together AI API. For the first approach, I utilized the BERT tokenizer and

pre-trained BERT model (bert-base-uncased) to generate contextual embeddings for each summary. The summaries were tokenized and converted into tensors, which were then passed through the BERT model. Embeddings were extracted from the model's last hidden state and averaged across all tokens to produce fixed-length vector representations of the summaries. These embeddings encapsulate the semantic meaning of the text and serve as a foundation for identifying key terms. Simultaneously, I applied Term Frequency-Inverse Document Frequency (TF-IDF) with n-gram support (unigrams, bigrams, and trigrams) to pinpoint the most significant terms within each summary. By ranking terms based on their importance, the top six keywords were identified and stored in a new column, keywords_using_bert. This data was saved in an intermediate file, cleaned_ml_papers_with_bert.csv. In the second approach, I used the Together AI API for keyword extraction. A custom prompt was designed to instruct the model to generate up to six keywords in a pipe-separated format for each summary. The model's responses were processed to extract these keywords, which were stored in a separate column, keywords_using_togetherai. The final output was a comprehensive dataset saved as cleaned_ml_papers_with_bert_and_togetherai.csv. This file contained the original summaries alongside two additional columns for keywords extracted using BERT embeddings with TF-IDF and the Together AI API. Finally, I manually curated set of keywords, which served as the ground truth for evaluating the performance of the keyword extraction methods throughout this study.

For the second task, I extended my analysis by using the same dataset of 200 research papers. Using the Together AI API, I designed a customized prompt instructing the AI to generate hypotheses based on the provided paper summaries. This allowed to evaluate the AI's ability to propose research ideas that align with the content of the papers. Additionally, I manually generated hypotheses for all papers based on my own understanding of the summaries. To assess the quality of both AI-generated and human-generated hypotheses, I employed the Together AI API to rate each hypothesis on several factors: novelty, excitement, feasibility, and effectiveness, using a scale from 0 to 10. The ratings were then averaged across all categories for each hypothesis to provide a comprehensive evaluation.

Third task was topic modeling. I utilized a multi-step approach to perform topic modeling on research papers obtained from the arXiv API. First, I fetched the papers using a Python script that queries the arXiv API with a specified search term. The fetch_arxiv_data function handles the retrieval of metadata including titles, authors, publication dates, summaries, and links to full papers, using the requests library to send the API request and BeautifulSoup for XML parsing. The fetched data is then stored in an SQLite database, where I created a table containing fields for paper titles, authors, publication dates, summaries, and links. The next step involved preprocessing the paper summaries: I cleaned the text by converting it to lowercase, tokenizing the content, and removing stopwords, which were then stored in a new column for further analysis. For topic modeling, I used Latent Dirichlet Allocation (LDA) to identify underlying topics within the summaries. A document-term matrix (DTM) was constructed using the CountVectorizer to convert the cleaned text into a matrix of word counts. The LDA model was applied to the DTM to extract five topics, and the top 20 words for each topic were analyzed to characterize the themes. Finally, I visualized the topic modeling results using pyLDAvis, which provided an interactive representation of the relationship between words and topics. The visualization was saved as an HTML file for exploration and interpretation of the topics.

3 Results

Keywords Evaluation: I used 0.7 as the standard cosine similarity threshold, meaning that if the cosine similarity exceeds this value, we can say that the model's performance is in line with human-annotated keywords. Based on this threshold, I compared the performance of BERT and Together AI. Together AI showed a higher proportion of threshold exceedance (0.83) compared to BERT (0.58) (Table 1) (0 when threshold not exceeded, 1 when exceeded), indicating a stronger alignment with human-annotated keywords. Additionally, Together AI achieved a higher mean cosine similarity score (0.8093) than BERT (0.6559), with lower variability, as evidenced by its smaller standard deviation (0.1342 vs. 0.1709 for BERT). Cohen's Kappa values also highlighted that Together AI had a stronger agreement with the ground truth (0.66) compared to BERT (0.17). Despite these

differences, the correlation between the cosine similarity values of BERT and Together AI was low (0.1238), suggesting that the models capture different aspects of the data. The bar graph clearly illus-

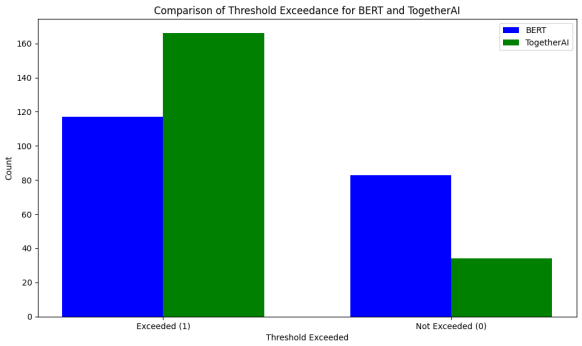

Figure 1: Cosine Similarity Threshold Exceedance Comparison

trates that BERT has a higher count of instances where the threshold was not exceeded compared to Together AI, indicating that BERT's performance aligns less frequently with the human-annotated keywords (Figure 1).

Metric	BERT	Together AI
Threshold Exceedance Counts	1: 117, 0: 83	1: 166, 0: 34
Mean Cosine Similarity	0.6559	0.8093
Median Cosine Similarity	0.7216	0.8247
Count of Threshold Exceedance (1's)	117	166
Proportion of Threshold Exceedance	0.58	0.83
Standard Deviation of Cosine Similarity	0.1709	0.1342
Agreements with Ground Truth	117	166
Disagreements with Ground Truth	83	34
Cohen's Kappa with Ground Truth	0.17	0.66

Table 1: Metrics Comparison between BERT and Together AI

Hypotheses Evaluation: In evaluating the hypotheses generated by both human and AI, I observed distinct trends across the four categories: novelty, excitement, feasibility, and effectiveness. AI outperformed humans in terms of novelty and excitement, with average scores of 8.88 and 9.22, respectively, compared to 8.28 and 8.75 for humans (Figure 2, 3).

On the other hand, humans excelled in feasibility and effectiveness (Figure 4, 5), with scores of 9.19 and 9.04, respectively, compared to 7.67 and 7.89 for AI. This indicates that human-generated hypotheses are more grounded in practical considerations and real-world applicability. When looking at the overall performance, the human-generated hypotheses proved to be more balanced, with a final average score of 8.81, while AI scored 8.41 overall (Table 2, Figure 6). Although AI showed a strong

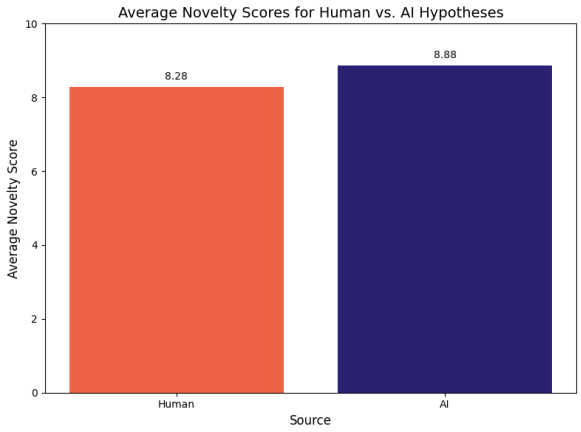

Figure 2: Novelty Scores

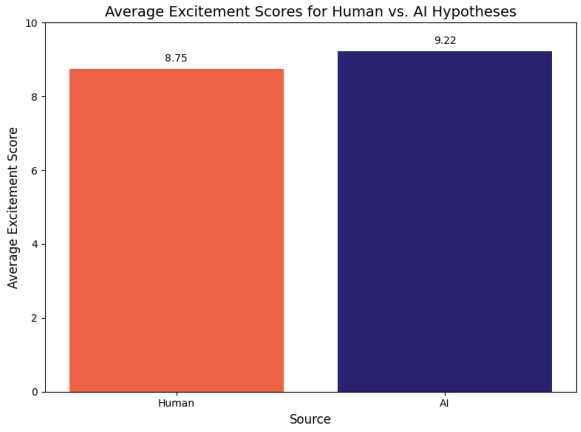

Figure 3: Excitement Scores

ability to generate innovative and exciting ideas, humans demonstrated superior ability in delivering feasible and effective solutions.

Category	Average Human Score	Average AI Score
Novelty	8.28	8.88
Excitement	8.75	9.22
Feasibility	9.19	7.67
Effectiveness	9.04	7.89
Overall	8.81	8.41

Table 2: Comparison of Average Human and AI Scores across Categories

Topic Modeling: The topic modeling analysis revealed several key thematic clusters within the corpus of NLP research papers. One prominent cluster centered around language and translation (Figure 7), as evidenced by the high-frequency terms "languages," "translation," "task," "natural," and "review." This suggests a strong emphasis on multilingual and cross-lingual NLP applications in the literature, reflecting ongoing efforts to improve language understanding across different linguistic

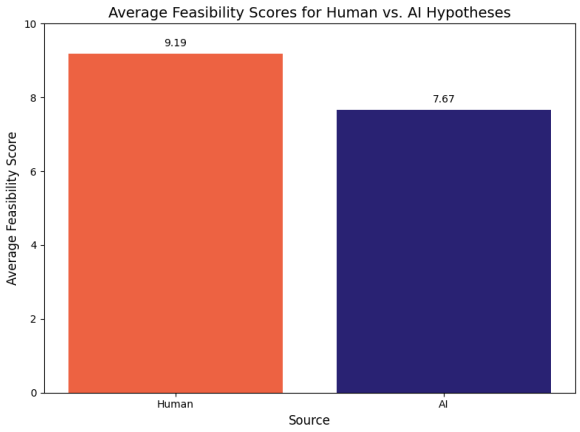

Figure 4: Feasibility Scores

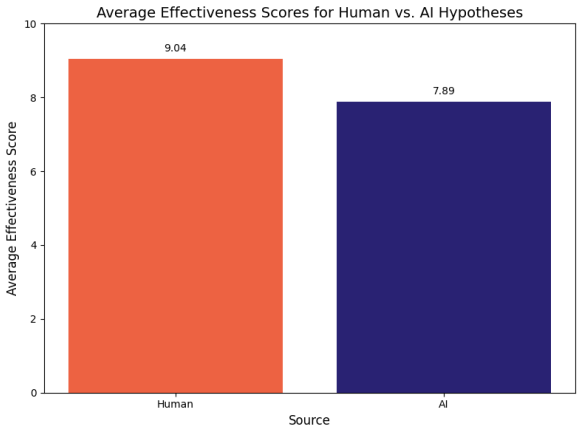

Figure 5: Effectiveness Scores

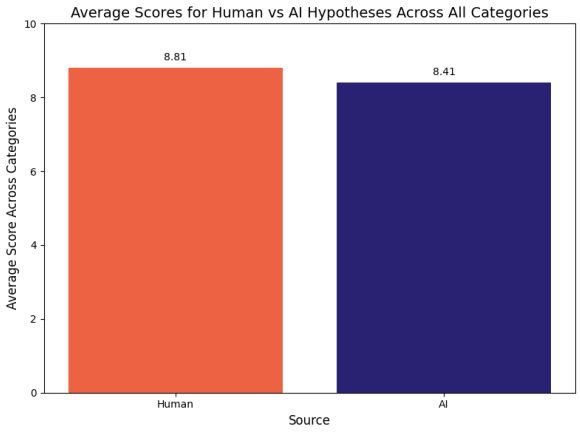

Figure 6: Overall Average

boundaries. Another distinct cluster focused on reasoning and graph-based approaches, highlighted by keywords such as "reasoning," "graph," "affinitize," and "brain." This indicates that researchers are increasingly exploring the use of graph-based models and reasoning techniques in NLP tasks, particularly for complex problem-solving and inference

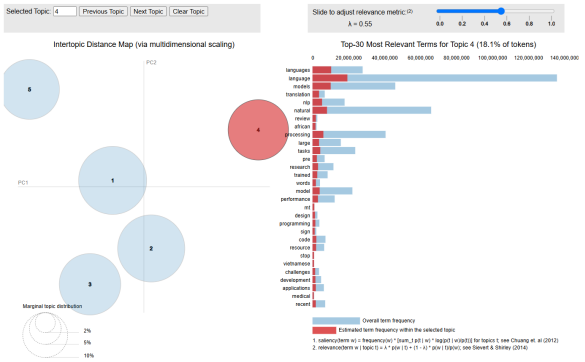

Figure 7: Top 5 Topics Identified Through LDA Topic Modeling

generation. A third cluster emerged around representation and modeling, characterized by terms like "knowledge," "embedding," "model," "quantum," "signals," and "mt." This points to significant advancements in knowledge representation and machine learning models, particularly their application in fields such as machine translation, where embedding techniques and advanced models are being leveraged to enhance system performance. Finally, a set of topics emerged around data-efficient and programmatic learning, signified by keywords like "low," "data," "programming," "logic," and "shot." This suggests an increasing focus on developing NLP systems capable of learning efficiently with limited data, as well as the exploration of programmatic and logic-based approaches to training and improving models.

4 Conclusion

This study explored the feasibility of using AI for three key aspects of academic paper writing: automatic keyword generation, hypothesis generation, and topic modeling. In terms of automatic keyword extraction, I found that the Together AI API, when used with a customized prompt, outperformed the transformer-based language model BERT. The customized prompt likely allowed the AI to better understand the nuances of the research area, leading to more accurate and contextually relevant keywords. Given these results, we can confidently conclude that AI is a viable tool for automatic keyword generation, offering an efficient and reliable method for enhancing research paper content.

When evaluating automatic hypothesis generation, a comparison between human-generated hypotheses and those generated by Together AI re-

vealed mixed results. AI performed better in terms of novelty and excitement, suggesting that it is capable of generating fresh and innovative ideas. However, in terms of feasibility and effectiveness, human-generated hypotheses scored higher, highlighting the challenge AI faces in producing hypotheses that are realistic and practical in specific research contexts. Despite these differences, the overall scores were close, with human-generated hypotheses scoring an average of 8.81 and AI-generated ones scoring 8.41. This suggests that while AI is capable of generating viable hypotheses, it is still essential to rely on human expertise for assessing feasibility and effectiveness.

For topic modeling, we identified the top five topics and their associated keywords, which appeared to align well with the key themes in the research papers. This indicates that AI-driven topic modeling can successfully discover and group relevant terms within a specific research domain. The ability to automatically identify key topics and trends in research papers makes topic modeling a valuable tool for academic writing. In conclusion, AI shows significant potential across various aspects of academic writing, including generating accurate keywords, identifying relevant topics, and proposing innovative hypotheses, thereby providing valuable support to the research process.