Exploratory Data Analysis and Modeling of Large Language Model Survey Papers

Guanming Zhang
Department of Computer Science
Boise State University
guanmingzhang@u.boisestate.edu

Abstract

In this report, we present an exploratory data analysis of a dataset containing survey papers related to large language models (LLMs). The analysis is performed using a combination of data manipulation techniques with Pandas, visualized with Matplotlib, and preprocessed using Sklearn for modeling purposes. Key insights regarding trends, popular research topics, and metadata of the surveyed papers are discussed. This analysis aims to provide a better understanding of the recent advancements and research directions in the field of LLMs, contributing to the broader study of artificial intelligence. Our report includes data exploration, modeling performance, and a discussion of the results, offering a clear methodology for analyzing academic datasets.

1 Introduction

AI techniques have been widely applied to various domains, such as images (He et al., 2016; Dosovitskiy, 2020), texts (Vaswani et al., 2017; Devlin et al., 2018), and graphs (Kipf and Welling, 2016; Zhuang and Al Hasan, 2022). As a critical subset of AI techniques, Large Language Models (LLMs) have gained significant attention in recent years (Radford et al., 2018, 2019; Brown et al., 2020; Achiam et al., 2023; Bai et al., 2022; Team et al., 2023). Especially, more and more new beginners are interested in the research topics about LLMs. To learn the recent progress in this field, new beginners commonly will read survey papers about LLMs. Therefore, to facilitate their learning, numerous survey papers on LLMs have been published in the last two years. However, a large amount of these survey papers can be overwhelming, making it challenging for new beginners to read them efficiently. To embrace this challenge, in this project, we aim to explore and analyze the metadata of LLMs survey papers, providing insights to enhance their accessibility and understanding (Zhuang and Kennington, 2024). Specifically, we aim to conduct a comprehensive exploratory data analysis of the metadata associated with LLM survey papers. Our analysis will focus on identifying key trends in LLM research, the most frequently covered topics, the distribution of paper categories, and the publication channels. We will employ several data processing techniques for data exploration visualizations, and analysis. By doing so, we hope to provide valuable insights that can guide new researchers in navigating the overwhelming volume of LLM literature. Furthermore, we aim to create models that evaluate the research impact of these papers, potentially identifying which areas of LLM research are gaining the most traction and which are less explored. The results of this analysis will ultimately contribute to a better understanding of the LLM research landscape.

Overall, our contributions can be summarized as follows:

- We conducted an in-depth exploratory data analysis (EDA) to identify key trends, popular topics, and emerging themes in LLM research.
- We visualized the distribution of paper categories, publication channels, and author collaborations using various graphical techniques, offering insights into the structure and dynamics of the LLM research community.
- We applied data preprocessing techniques to prepare the metadata for analysis, ensuring data quality and consistency.
- We documented all findings and methodologies in a clear, structured report using LaTeX, adhering to academic standards for clarity and reproducibility.

2 Related Work

Large Language Models (LLMs) have been widely explored in recent years, with significant advancements in architectures, such as Transformer models (Vaswani et al., 2017), and their applications across various domains, including natural language

processing (NLP) and artificial intelligence (Radford et al., 2019; Brown et al., 2020). Several survey papers have provided comprehensive overviews of LLMs, focusing on topics like model performance, ethical considerations, and practical applications (**?**Zhuang and Kennington, 2024). However, few studies have specifically focused on analyzing the metadata of LLM survey papers. This work aims to fill that gap by conducting an exploratory analysis of LLM survey paper metadata to uncover trends and insights in this rapidly evolving field.

3 Methodology

3.1 Data Exploration

In this section, we performed feature extraction and data preprocessing steps to prepare the data for machine learning model training and evaluation. Using the Pandas, sklearn, and scipy libraries, we conducted tasks such as feature extraction, data normalization, label encoding, and dataset splitting. First, we combined the Title and Summary columns and applied TF-IDF (Term Frequency-Inverse Document Frequency) vectorization using the TfidfVectorizer. This allowed us to extract important words from each document and represent them as numerical features in a matrix. To handle the categorical Categories column, we applied One-Hot Encoding to convert the categories into a numerical format. Finally, we used the hstack function to horizontally combine the TF-IDF and One-Hot encoded features, creating the final feature matrix. Next, we carried out further preprocessing on the feature matrix. We normalized the features using MinMaxScaler, which scales all feature values to a range between [0,1], ensuring that features of different magnitudes do not disproportionately influence the model. We then used LabelEncoder to encode the target labels, transforming categorical labels into numeric values. Finally, we split the dataset into training and testing sets using the train-test-split function, with 40 percent of the data allocated to the test set and 60 percent to the training set. This split ensures that the model is trained and evaluated on separate data, enabling reliable performance assessment.

3.2 Data Manipulation

In this section, we performed feature extraction and data preprocessing steps to prepare the data for machine learning model training and evaluation.

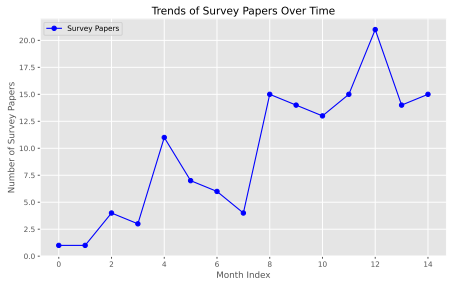

Figure 1: Trends of Survey Papers Over Time

Using the Pandas, sklearn, and scipy libraries, we conducted tasks such as feature extraction, data normalization, label encoding, and dataset splitting.

First, we combined the Title and Summary columns and applied TF-IDF (Term Frequency-Inverse Document Frequency) vectorization using the TfidfVectorizer. This allowed us to extract important words from each document and represent them as numerical features in a matrix. To handle the categorical Categories column, we applied One-Hot Encoding to convert the categories into a numerical format. Finally, we used the hstack function to horizontally combine the TF-IDF and One-Hot encoded features, creating the final feature matrix.

Next, we carried out further preprocessing on the feature matrix. We normalized the features using MinMaxScaler, which scales all feature values to a range between [0,1], ensuring that features of different magnitudes do not disproportionately influence the model. We then used Label Encoder to encode the target labels, transforming categorical labels into numeric values. Finally, we split the dataset into training and testing sets using the train_test_split function, with 40% of the data allocated to the test set and 60% to the training set. This split ensures that the model is trained and evaluated on separate data, enabling reliable performance assessment.

These data manipulation steps successfully produced a feature matrix ready for training, ensuring both data quality and consistency throughout the process.

3.3 Data Evaluation

To evaluate the performance of the machine learning model, we employed a Logistic Regression model to classify the documents based on their features. The model was trained using an 80% training set and evaluated on a 20% test set. Several metrics,

including accuracy, F1-score, and a detailed classification report, were used to assess the model's performance.

The initial evaluation of the Logistic Regression model without any class balancing yielded an accuracy of 0.5517. However, the classification report showed varying performance across different classes, with some classes achieving high precision and recall (e.g., "RecSys IR" and "Robotics") while others had poor performance (e.g., "Adaptation Tuning" and "Evaluation"). This suggests that the dataset suffers from class imbalance, where certain categories have fewer instances, leading to reduced predictive performance in those categories.

To mitigate this issue, we applied class balancing using the class_weight='balanced' parameter in Logistic Regression, which adjusts the weights inversely proportional to the class frequencies. This approach improved the F1-score for several underrepresented classes. After re-training the model with class balancing, the weighted F1-score improved to 0.5000, although accuracy remained relatively low.

The detailed classification report revealed insights into how the model performed across different categories. For instance, categories such as "RecSys IR" and "Robotics" had perfect precision and recall, indicating that the model was very accurate in predicting these categories. However, other categories like "Adaptation Tuning" and "Multi-modal Pre-training" had significantly lower performance, highlighting areas where the model struggled to generalize.

In conclusion, the evaluation demonstrated that while the Logistic Regression model performed reasonably well in certain categories, class imbalance and the limited number of samples in some categories negatively impacted its overall performance. Addressing these issues, perhaps through more advanced techniques such as data augmentation or using more complex models, may help improve the classification performance further.

Title 1	Title 2
XX	XX
XX	XX

Table 1: TODO: Describe this table.

4 Conclusion

A APPENDIX

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Alexey Dosovitskiy. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jun Zhuang and Mohammad Al Hasan. 2022. Defending graph convolutional networks against dynamic graph perturbations via bayesian self-supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 4405–4413.

Jun Zhuang and Casey Kennington. 2024. Understanding survey paper taxonomy about large language models via graph representation learning. *arXiv preprint arXiv:2402.10409*.