Exploring Large Language Models Survey papers using Random Forest Classification with Over Sampling the Minority classes

Anh Bui
Department of Computer Science
Boise State University
anhbui@u.boisestate.edu

Abstract

The rapid advancement in Large Language Models (LLMs) has made it challenging for researchers to keep up with new models and innovations. While many scholars have published survey papers to synthesize this work, the growing number of surveys has added complexity, making it harder to stay updated on recent developments. In this report, I present an analysis of a dataset comprising 144 survey papers published between 2023 and January 2024. To address the dataset's class imbalance, I applied oversampling techniques, specifically the Synthetic Minority Over-sampling Technique (SMOTE) and Random Over Sampling. These methods, combined with a Random Forest Classifier, were used to predict the 'taxonomy class' of new papers, resulting in an improvement in the model's precision score from 0.35 to 0.42. Future work will focus on enhancing classification accuracy and exploring other machine learning algorithms to expand the applicability of this approach.

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, I aim to explore and analyze the metadata of LLMs survey papers, providing insights to enhance their accessibility and understanding (Zhuang and Kennington, 2024). Specifically, I aim to build a report that will provide key trends over time and the distribution of taxonomy classes, categories within the dataset. Additionally, the goal is to develop the robust model capable of classifying new survey papers into the appropriate taxonomy classes. This classification will assist readers, particularly beginners, in selecting survey papers based on their specific needs and research objectives.

Overall, my contributions can be summarized as follows:

- I analyzed the data set of 144 survey papers and their metadata to understand this characteristics, such as: the trend of survey paper over time, the distribution of paper's category and also the distribution of the paper's taxonomy.
- I conducted and processed a feature matrix using the paper titles, summaries and categories to develop the model for classifying the taxonomy of new papers.
- I applied Random Forest Classification with Over Sampling the Minority Class (using Synthetic Minority Over-sampling Technique(SMOTE) and Random Over Sampling) to address the imbalance issue of this data set, which improving the precision score of the model from 0.35 to 0.42

2 Related Work

3 Methodology

3.1 Data Exploration

The primary goal of this step is to gain a thorough understanding of the dataset and address any data

quality issues, such as duplicates and missing values, in preparation for the subsequent data manipulation steps

The first task was to examine the temporal distribution of the number of published papers per month. To achieve this, the "Release Date" field was converted into a datetime format to allow for accurate time-based analysis. The data was then grouped by month, and a line chart was used to visualize the trend of papers published over time, as shown in Figure 1. Next, I explored the distribution of taxonomy classes within the dataset, as this factor plays a critical role in determining the approach for building a model to classify survey papers in later steps. To assess this distribution, I counted the number of instances for each taxonomy class and visualized the results using a bar chart, which is presented in Figure 2. The chart revealed that the dataset is imbalanced, with certain taxonomy classes being underrepresented, such as "Finance", "Education", and "Hardware Architecture", each containing only one sample while the "trustworthy" class has 26 samples. This imbalance will need to be addressed during the data manipulation and model-building phases to ensure accurate classification

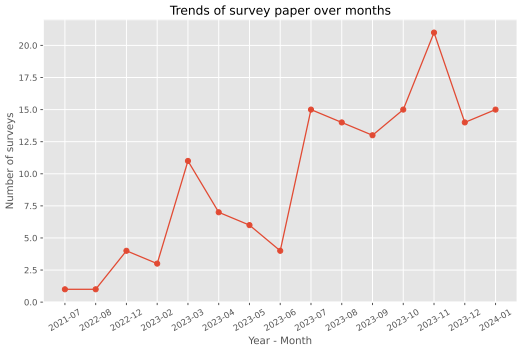

Figure 1: The number of published survey paper over months

Additionally, I examined the distribution of the "categories" to which the survey papers belong, as classified by ArXiv. To assess this distribution, I split each category into a new column, counted the number of instances for each category, and visualized the results using a bar chart, which is presented in Figure 3.

3.2 Data Manipulation

In this section, I constructed and preprocessed a feature matrix. Text-based features from the "Title" and "Summary" columns were vectorized using the Term Frequency-Inverse Document Frequency (TF-IDF) method, transforming them into numerical values. Since several cells in the "Categories" column contained multiple values (e.g., "cs.AI, cs.ML, cs.LL"), I employed the MultiLabelBinarizer for one-hot encoding, enabling better handling of multi-category entries.

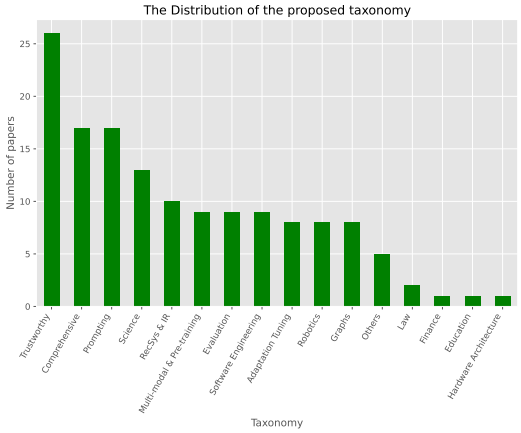

Figure 2: The distribution of classes in taxonomy

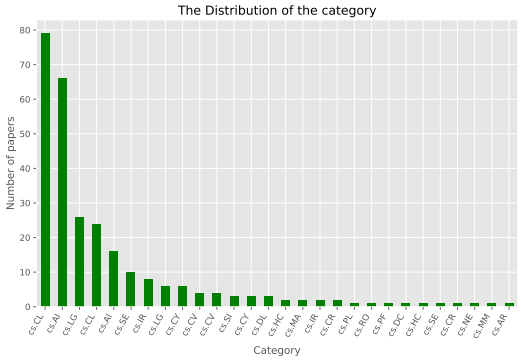

Figure 3: Distribution of survey papers that I found across different arXiv categories

The TF-IDF and one-hot encoded matrices were then concatenated to form a comprehensive feature matrix. I normalized the data using the MinMaxScaler to ensure all features were on a consistent scale and encoded the "Taxonomy" labels as numerical values. Finally, the dataset was split into training and testing sets using a 60/40 ratio to facilitate model evaluation.

3.3 Data Evaluation

Given the imbalance in the dataset, I selected Random Forest Classification as the model and used the precision score as the primary evaluation metric. Initially, using the default parameters of Random Forest Classification, the precision score was 0.35.

In a second attempt, I addressed the imbalance is-

Attempt	Precision-score
Attempt 1	0.35
Attempt 2	0.42

Table 1: Model precision-score comparision.

sue by applying Oversampling the Minority Class, using the Synthetic Minority Over-sampling Technique (SMOTE) and Random Over Sampling. Additionally, I adjusted the class weight parameter in the Random Forest model. This resulted in an improved precision score of 0.42.

4 Conclusion

In this project, I analyzed a dataset of 144 LLM survey papers, identifying key trends and imbalances in taxonomy and categories. After preprocessing the data using TF-IDF vectorization and one-hot encoding, I applied Random Forest Classification. To address the dataset's imbalance, I employed Random Over Sampling and SMOTE, which, combined with adjusted class weights, improved the precision score from 0.35 to 0.42. These results highlight the importance of addressing data imbalance for improving classification performance.

In future work, I will focus on updating the dataset and exploring new methods to further enhance the model.

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*.