LLM Survey Analysis Using Random Forest

Priyanka Singla
Department of Computer Science
Boise State University
priyankasingla@u.boisestate.edu

Abstract

This project investigates the application of a Random Forest Classifier for analyzing metadata from survey papers on large language models (LLMs), a rapidly growing area within AI. The goal is to assist new researchers by providing insights into the trends and patterns in LLM survey publications. Through a structured workflow—comprising data loading, exploration, manipulation, and visualization—key attributes such as release dates, categories, and taxonomies were analyzed. Techniques like TF-IDF vectorization, one-hot encoding, and feature scaling were employed to construct a robust feature matrix. Hyperparameter tuning using grid search optimized the classifier's performance. Although the model achieved perfect training accuracy, a lower test accuracy (0.39) indicated overfitting, likely caused by dataset imbalance. With a best cross-validation score of 0.26, future improvements will focus on addressing data imbalance, enhancing feature engineering, and exploring alternative models to boost performance. The project highlights trends in LLM research and suggests paths for enhancing model accuracy.

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).

Proposed Plan:

Specifically, My objective is to systematically analyze the dataset through data loading, exploration, manipulation, and visualization, with the aim of extracting meaningful insights and enhancing both accessibility and understanding.

The methodology is divided into the following key phases: Data Loading, Data Exploration, Data Visualization, and Data Validation. Each phase is explained in detail in the later sections of this report.

Based on this approach, the timeline for the project is structured as follows:

Week 1: Data loading and initial exploration.

Week 2: Data manipulation and cleaning.

Week 3: Visualization and model validation.

Week 4: Finalizing the report in LaTeX.

To achieve optimal results, I utilized several libraries, including Matplotlib and Seaborn, among others, for data visualization and analysis.

In short, this proposed plan outlines a systematic approach for analyzing the dataset, allowing for the extraction of meaningful insights and the validation of findings through model testing.

Overall, My contributions to the project can be summarized as follows:

a) Implementation of a structured data analysis workflow.

b) Effective use of visualization libraries for extracting insights.

c) Generated bar charts and box plots to effectively findings and highlighting patterns in the data.

d) Thorough validation of findings through model testing and analysis, then applied the Random Forest algorithm to evaluate accuracy.

2 Methodology

2.1 Data Exploration

This section presents a overview of the key findings from the data exploration process:

a) Dataset Overview: The dataset comprises multiple attributes, including 'Release Date', 'Title', 'Summary', 'Categories', and 'Taxonomy', which are essential for understanding survey papers.

b) Data Types: The 'Release Date' is formatted as a datetime object, while 'Categories' and 'Taxonomy' are categorical variables, which are crucial for subsequent analyses.

c) Missing Values: The dataset was examined for missing values, revealing that all critical columns are complete, ensuring data integrity for analysis.

d) Descriptive Statistics: The average number of surveys released each month is approximately 9.6, providing a baseline for understanding publication trends.

e) Trends Over Time: By grouping data by year and month, we identified fluctuations in survey releases, highlighting periods of increased or decreased publication activity. (Fig. 1)

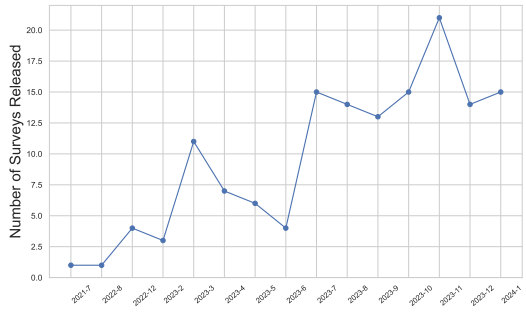

Figure 1: Trends of survey papers over month

f) Taxonomy Distribution: Analysis of the proposed taxonomy revealed that 26 survey papers are categorized as "Trustworthy," indicating a significant focus on this area within the dataset. (Fig. 2)

g) Visual Insights: Line Plot: Depicted trends in survey releases over time, showcasing publication patterns. (Fig. 1)

Bar Chart: Illustrated the distribution of survey papers across various taxonomy categories, clarifying research focus areas. (Fig. 2)

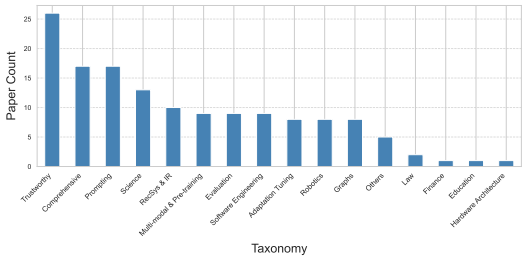

Figure 2: Distribution of the proposed taxonomy

Box Plot: Analyzed release days, providing insights into publication timing relative to different Taxonomy. (Fig. 3)

In addition to that I used Box Plot for the above analysis and visualise the dataset. Box plots are considered efficient data visualization tools because they provide a concise summary of a dataset's distribution in a single, compact graphic. They are particularly useful for comparing multiple datasets or groups and it shows the distribution of release days for each taxonomy. Each box represents a taxonomy, and the components of the box plot provide the following information:

The box plot visualizes the distribution of paper release days across AI subfields. Each box represents the interquartile range (IQR), containing the middle 50% of the data, with the line inside showing the median (50th percentile), or typical release day. The whiskers extend beyond the IQR, showing the overall range of release days, while individual points outside the whiskers represent outliers—unusually early or late releases.

Tall boxes, such as in "Multi-modal," indicate greater variability in release times, suggesting papers are published more irregularly. Shorter boxes, like in "Law," imply more consistent timing. Outliers, particularly in fields like "Pretraining," might indicate groundbreaking papers released outside usual patterns.

Comparing median lines across fields (e.g., "Graphs" vs. "RecSys & IR") shows which categories have papers released earlier or later. Skewness in the boxes can highlight whether a field tends to have earlier or later paper releases, helping to understand trends and activity timing across AI research.

2.2 Data Manipulation

In Data Manipulation outlines the steps taken to create a feature matrix from the dataset:

a) Vectorization of Text Data: The 'Title' and

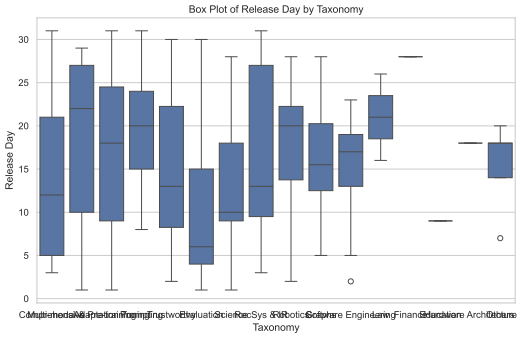

Figure 3: Box Plot

'Summary' columns were converted into numerical representations using TF-IDF vectorization with TfidfVectorizer, creating a sparse matrix that captures the importance of words while excluding common stop words.

b) One-Hot Encoding for Categorical Data: The 'Categories' column was split into separate entries and one-hot encoded using pd.getdummies, transforming categorical values into a binary format suitable for modeling.

c) Combining Features: A comprehensive feature matrix was constructed by combining the TF-IDF matrices for 'Title' and 'Summary' with the one-hot encoded categories.

d) Normalization of Features: The feature matrix was normalized using MinMaxScaler to scale all features to a range between 0 and 1, which is essential for many machine learning algorithms.

e) Label Encoding: The target variable, 'Taxonomy', was encoded using LabelEncoder, converting categorical labels into numerical values for model training.

f) Dataset Splitting: The dataset was split into training and testing sets (70/30) using train and test split, allowing for effective model training and evaluation. These steps ensured that the raw dataset was effectively transformed into a structured feature matrix, ready for analysis and modeling.

2.3 Data Evaluation

I employed the Random Forest Classifier to evaluate my dataset. This model is highly appropriate for the task, offering accurate predictions while capturing intricate relationships and emphasizing key features. Its flexibility and robust performance make it especially capable of managing varied and noisy data. Furthermore, I applied grid search for hyperparameter optimization to improve accuracy. For data evaluation, the sklearn library was used to train the model and test its accuracy.

3 Conclusion

In conclusion, I applied a Random Forest Classifier, fine-tuned through hyperparameter optimization, with the following best parameters:

Max Depth: 15

Min Samples Leaf: 1

Min Samples Split: 2

Number of Estimators: 200

The model achieved a perfect training accuracy of 1.0 but only 0.39 on the test data, indicating overfitting. This discrepancy is likely due to an imbalanced dataset, primarily composed of 0 and false values. The best cross-validation score was 0.26, suggesting further enhancements are needed. Future work should focus on addressing the dataset imbalance, applying more feature engineering, and considering alternative modeling techniques to improve performance.

A APPENDIX

* This section summarizes the key settings and hyperparameters used for the Random Forest Classifier in this analysis.

1. Model Type:

Classifier: Random Forest Classifier

2. Hyperparameters:

Number of Estimators: 200

Max Depth: 15

Min Samples Split: 2

Min Samples Leaf: 1

3. Training and Testing Configuration:

Data Split: 70% for training and 30% for testing

Cross-Validation: 5-fold

4. Data Preprocessing:

Vectorization: Applied TF-IDF to the 'Title' and 'Summary'

Normalization: Used Min-Max scaling on features

Label Encoding: Encoded the target variable 'Taxonomy' with LabelEncoder

These settings were crucial for enhancing the model's performance and improving the accuracy of predictions.

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*.