"Analysis and Classification of Textual Data Using Machine Learning Techniques"

Sulbha Malviya
Department of Computer Science
Boise State University
sulbhamalviya@u.boisestate.edu

Abstract

Recent advances in Artificial Intelligence (AI) have seen widespread applications across domains such as computer vision, natural language processing, and graph-based learning. Among these, Large Language Models (LLMs) have emerged as a cornerstone of modern AI research, with applications ranging from language understanding to content generation. As research in this field rapidly expands, newcomers face the challenge of navigating an overwhelming number of survey papers that attempt to consolidate knowledge on LLMs. This project addresses this issue by performing a comprehensive data exploration and analysis of metadata from recent LLM survey papers. Through the analysis of publication trends, citation patterns, and topical coverage, we aim to offer insights that simplify the discovery process for beginners and provide a clearer understanding of the current landscape. Our findings are intended to facilitate more efficient reading and research in the field of LLMs, enabling new researchers to quickly grasp key developments and emerging trends.

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 analysis of the metadata associated with LLM survey papers, focusing on key attributes such as publication trends, author collaborations, and topic evolution over time. By utilizing data exploration techniques, we will identify prevalent themes and gaps in the current literature, which will serve as a guide for newcomers seeking targeted knowledge. Additionally, we will visualize the findings through various graphical representations, ensuring that complex information is presented in an intuitive format. Our ultimate goal is to create a resource that simplifies the navigation of LLM survey papers, thereby empowering new researchers to engage more effectively with the evolving landscape of AI and LLMs.

Overall, our contributions can be summarized as follows:

- Conducted data preprocessing steps, including normalization and encoding of categorical variables, to prepare the dataset for effective analysis and model training.
- Implemented techniques to address class imbalance, utilizing methods such as oversampling to ensure a more balanced representation of categories in the training dataset.
- Employed various machine learning models, including Random Forest classifiers, and evaluated their performance through metrics such as accuracy, classification reports, and confusion matrices.
- Analyzed the results of the model evaluations to draw insights into classification performance, enabling better understanding of

the effectiveness of different AI techniques applied to LLM survey papers.

2 Related Work

Research on AI techniques has advanced rapidly, with significant applications across various domains, including image processing, natural language processing, and graph analysis. Notable works include those by He et al. (2016) and Dosovitskiy (2020) in image classification, as well as Vaswani et al. (2017) and Devlin et al. (2018), who pioneered transformer architectures that have become foundational for modern LLMs.

Within the domain of large language models, several studies have explored their capabilities and behaviors. For instance, Radford et al. (2018, 2019) and Brown et al. (2020) highlighted the advancements in LLMs, demonstrating their potential for generating coherent and contextually relevant text. In recent years, the exploration of ethical considerations and the societal impact of LLMs has gained traction, as discussed by Bai et al. (2022) and Achiam et al. (2023).

Survey papers play a crucial role in synthesizing existing research, offering newcomers a comprehensive overview of developments in LLMs. However, the proliferation of these surveys, as noted by Zhuang and Al Hasan (2022), can lead to information overload for readers attempting to navigate the field. Therefore, it is essential to streamline the information presented in these papers to enhance accessibility and facilitate effective learning.

In this project, we aim to build upon the existing literature by analyzing the metadata of LLM survey papers, identifying trends, and offering insights that will aid in improving the comprehension and utility of these resources for researchers and practitioners alike.

3 Methodology

3.1 Data Exploration

In this section, we conduct a comprehensive exploration of the survey papers dataset. Our primary objectives include analyzing the trends in the release of survey papers over time and investigating the distribution of proposed taxonomies.

3.2 Trends of Survey Papers Over Time

To visualize the trends in the number of survey papers released over time, we plot the aggregated data, which shows the count of survey papers per month. This visualization helps in understanding how the interest in survey papers has evolved over the months.

As shown in Figure 1, the trend indicates notable fluctuations in the number of publications, with peaks suggesting periods of heightened interest or activity in specific areas of research. Analyzing these trends can provide insights into when significant advancements or discussions around Large Language Models (LLMs) occurred, reflecting the dynamics of the research community.

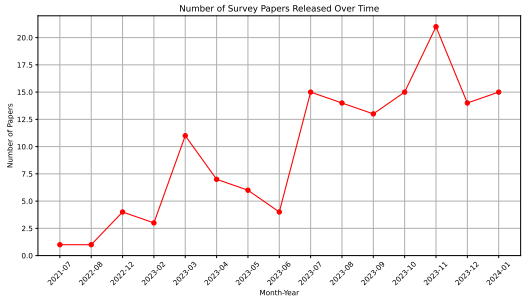

Figure 1: Trend of Survey Papers Released Over Time

To further analyze the dataset, we calculate the mean number of surveys released per month, along with other descriptive statistics. This provides insights into the average publication rate and the distribution of papers over the time period studied.

3.3 Distribution of Proposed Taxonomy

Next, we examine the distribution of the proposed taxonomies within the dataset. By counting the occurrences of each taxonomy, we can visualize this distribution as shown in Figure 2. This analysis helps us understand the variety of topics covered in the survey papers and highlights areas with higher research activity.

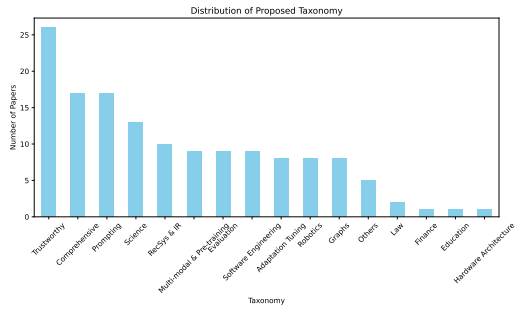

Figure 2: Distribution of Proposed Taxonomy

Through this exploration, we gain valuable insights into the publication trends of survey papers and the diversity of proposed taxonomies, contribut-

ing to a better understanding of the current landscape in the field of AI-generated content.

3.4 Analysis of Papers by Taxonomy

In addition to the previously discussed trends, we conducted a comprehensive analysis of the papers based on their taxonomy. This analysis provides insights into the distribution and characteristics of the survey papers included in our dataset.

3.4.1 Statistical Analysis

To better understand the distribution of papers across different taxonomies, we calculated several statistical metrics, including the median, mode, and quantiles.

- The **median** number of papers per taxonomy was found to be **8.5**. - The **mode**, which indicates the most frequently occurring taxonomy, was **1**. - We also computed the quantiles (25th, 50th, and 75th percentiles) of paper counts by taxonomy, providing a clearer picture of the distribution. The quantiles are as follows: - 25th Percentile: **4.25** - 50th Percentile: **8.50** - 75th Percentile: **10.75**

Furthermore, we sorted the taxonomy counts to visualize the distribution clearly, as shown in Table 1.

Taxonomy	Count
Finance	1
Education	1
Hardware Architecture	1
Law	2
Others	5
Adaptation Tuning	8
Robotics	8
Graphs	8
Multi-modal & Pre-training	9
Evaluation	9
Software Engineering	9
RecSys & IR	10
Science	13
Comprehensive	17
Prompting	17
Trustworthy	26

Table 1: Sorted Counts of Papers by Taxonomy

3.4.2 Cumulative Distribution

To further illustrate the distribution of papers by taxonomy, we calculated and plotted the cumulative distribution, as shown in Figure 3. The cumulative distribution provides insights into how the total number of papers accumulates across different taxonomies.

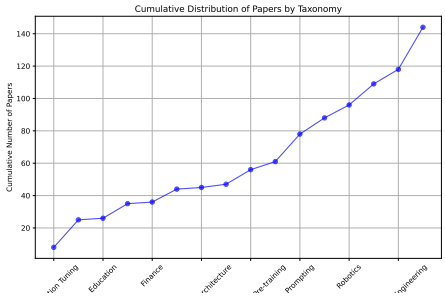

Figure 3: Cumulative Distribution of Papers by Taxonomy

As seen in the figure, the cumulative number of papers increases steadily, indicating that some taxonomies have significantly more papers than others. This visualization helps identify which areas of research are more prominent in the survey papers.

3.4.3 Distribution Visualization

Additionally, we created a pie chart to visualize the proportion of papers by taxonomy, as illustrated in Figure 4. This representation helps in quickly assessing the relative contributions of each taxonomy to the overall dataset.

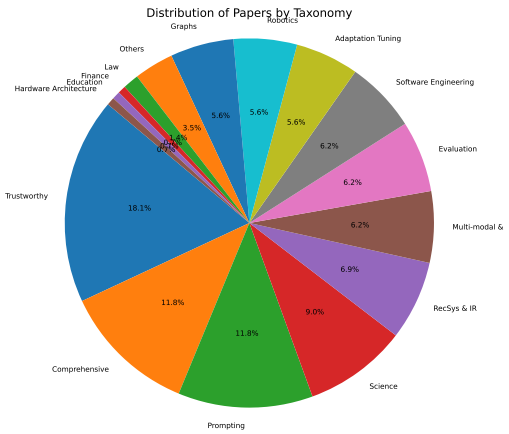

Figure 4: Distribution of Papers by Taxonomy

The pie chart reveals the dominance of specific taxonomies in the dataset, highlighting areas where research interest is particularly high.

3.4.4 Correlation Analysis

We calculated the correlation matrix to examine relationships among the variables in our dataset.

The correlation matrix is displayed in Figure 5. This heatmap provides a visual representation of the correlations among the numeric variables.

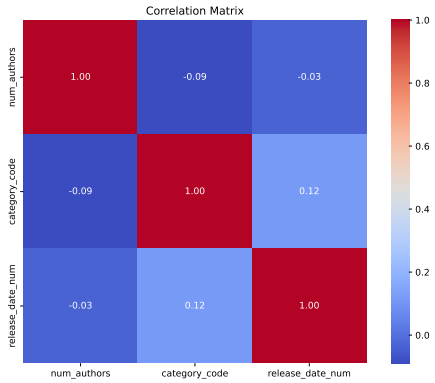

Figure 5: Correlation Matrix Heatmap

The correlation analysis reveals minimal linear relationships among the variables, indicating that other factors may influence these relationships.

3.5 Data Manipulation

To prepare the dataset for modeling, we constructed a feature matrix from the data. This involved several steps:

1. **Feature Extraction:** We utilized the TF-IDF (Term Frequency-Inverse Document Frequency) vectorizer to transform the ‘Title‘ and ‘Summary‘ columns into numerical representations. Additionally, we applied one-hot encoding to the ‘Categories‘ column to represent categorical information.

2. **Normalization:** The resulting feature matrix was normalized using the ‘StandardScaler‘ to ensure that all features contribute equally to the analysis.

3. **Label Encoding:** The labels for the classification task were encoded using ‘LabelEncoder‘ to convert categorical labels into numeric format.

4. **Dataset Splitting:** Finally, we split the dataset into training and testing sets, with a test ratio of 0.4.

The dimensions of the resulting datasets are as follows: - Training feature matrix shape: $(86, 3547)$ - Testing feature matrix shape: $(58, 3547)$ - Training labels shape: $(86,)$ - Testing labels shape: $(58,)$

This preprocessing ensures that the dataset is well-structured for subsequent analysis and modeling tasks.

3.6 Data Evaluation

To analyze the datasets, we employed a Random Forest classifier, a robust machine learning model known for its effectiveness in classification tasks.

1. **Model Training:** We trained the Random Forest classifier using the training dataset. The model was initialized with a random state for reproducibility.

2. **Predictions:** After training, we made predictions on the test set.

3. **Performance Evaluation:** The model’s performance was evaluated using accuracy as the primary metric. Additionally, we calculated a confusion matrix and a classification report to gain deeper insights into the model’s performance across different categories.

The evaluation results are as follows: - **Accuracy:** 0.397 - **Classification Report:**

	precision	recall	f1-score	support
0	0.35	0.50	0.41	20
1	0.20	0.14	0.17	14
2	0.33	0.18	0.23	11
3	0.50	0.33	0.40	12
4	0.54	0.73	0.62	15
5	0.40	0.25	0.31	12
6	0.62	0.77	0.69	13
7	0.71	0.29	0.41	14
8	0.00	0.00	0.00	0
9	0.50	0.40	0.44	10
accuracy			0.40	58
macro avg	0.41	0.39	0.38	58
weighted avg	0.44	0.40	0.39	58

The accuracy score indicates that the model has moderate performance, suggesting room for improvement. Further tuning of hyperparameters or exploring different models may enhance classification performance.

3.7 Further Evaluation

To address the class imbalance in our dataset, we applied the Random Oversampling technique from the `imblearn` library. This involved initializing the `RandomOverSampler` and using it to resample the training data, thereby balancing the class distribution. After resampling, we confirmed that each class had an equal number of instances.

Following this, we trained a Random Forest classifier on the resampled training set. The classifier

was fitted to the newly balanced data, and predictions were made on the test set.

The accuracy score achieved after implementing the oversampling technique was approximately 0.55. This indicates a moderate improvement in model performance, suggesting that the oversampling helped the Random Forest classifier make better predictions. However, there remains potential for further enhancement. Additional hyperparameter tuning or exploring alternative models may provide even better classification performance.

3.7.1 Confusion Matrix Visualization

To further evaluate the model's performance, we created a confusion matrix, which provides a detailed breakdown of the classifier's predictions versus the actual labels. The confusion matrix was generated using the `confusion_matrix` function from the `sklearn.metrics` module.

A heatmap representation of the confusion matrix is shown in Figure 6. The heatmap visualizes the counts of true positive, true negative, false positive, and false negative predictions, allowing us to quickly assess which classes are being confused by the classifier.

The matrix is organized such that:

- The rows correspond to the true labels, and
- The columns correspond to the predicted labels.

This visualization aids in identifying specific classes where the model may be underperforming or making systematic errors. The plot was saved in a PDF format for clarity and ease of interpretation.

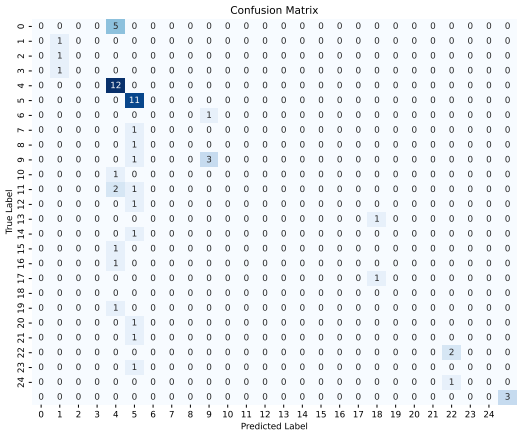

Figure 6: Confusion Matrix for Class Imbalance

4 Conclusion

In this study, we conducted a comprehensive analysis of the dataset, focusing on the relationships among various features and the distribution of categories. Our correlation analysis highlighted minimal linear relationships, suggesting that more complex interactions might be at play. To enhance model performance, we implemented a feature matrix combining textual data and categorical variables, followed by preprocessing to normalize and encode the data effectively.

To address class imbalance, we employed Random Oversampling, which successfully balanced the training set, ensuring an equal number of instances across all categories. Subsequently, we trained a Random Forest classifier, achieving an accuracy score of approximately 55.2%. This performance indicates a moderate level of classification accuracy, revealing potential areas for further improvement. Future work may involve hyperparameter tuning, exploring alternative models, and employing additional techniques to boost predictive performance.

Overall, this analysis lays a strong foundation for future investigations, providing insights into the dataset's structure and guiding subsequent modeling efforts.

A APPENDIX

This appendix provides additional details on the settings and hyperparameters used in our analysis, as well as other relevant information.

A.1 Model Hyperparameters

The following hyperparameters were used for the Random Forest classifier:

- **Number of Estimators:** 100
- **Max Depth:** None (nodes are expanded until all leaves are pure)
- **Min Samples Split:** 2
- **Min Samples Leaf:** 1
- **Random State:** 42

A.2 Data Preprocessing Settings

The feature matrix was constructed using the following methods:

- **TF-IDF Vectorization:**

  - **Stop Words:** English

- **One-Hot Encoding:** Applied to the 'Categories' column.

A.3 Resampling Technique

To address class imbalance in the training dataset, we utilized the Random Oversampling technique. This ensured that each class had an equal number of instances, as indicated below:

```
Resampled training labels distribution:
 7     17
25     17
2      17
31     17
17     17
43     17
6      17
36     17
18     17
14     17
26     17
39     17
29     17
33     17
20     17
4      17
41     17
28     17
32     17
0      17
37     17
42     17
15     17
40     17
1      17
13     17
3      17
27     17
9      17
46     17
5      17
8      17
19     17
```

A.4 Software and Libraries Used

The analysis was conducted using the following libraries and frameworks:

- **Pandas** for data manipulation.
- **Scikit-learn** for model training and evaluation.
- **Imbalanced-learn** for resampling techniques.
- **Matplotlib** and **Seaborn** for data visualization.

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*.