Automated Classification and Trend Analysis of Large Language Model Survey Papers Using Machine Learning and Natural Language Processing Techniques

Meherunnesa Tania
Department of Computer Science
Boise State University
meherunnesatania@u.boisestate.edu

Abstract

This study investigates the application of machine learning (ML) and natural language processing (NLP) techniques to classify academic survey papers into predefined taxonomy categories. The dataset, consisting of paper titles, summaries, release dates, taxonomy labels, and categories, was analyzed to uncover trends and patterns in the publication of research papers. Exploratory data analysis (EDA) revealed important insights through visualizations, such as publication trends over time, the distribution of taxonomy categories, and the most common terms used in paper summaries. Key NLP techniques, including Term Frequency-Inverse Document Frequency (TF-IDF), were employed to transform the textual data into numerical features, while one-hot encoding was applied to the categorical data. A Random Forest Classifier was trained on the extracted feature matrix to predict the taxonomy category of each paper. The model achieved promising accuracy, effectively capturing patterns in the dataset. The study also identified areas for future improvement, including addressing class imbalance and exploring more sophisticated models. These findings demonstrate the potential of ML and NLP for automating the classification of academic papers, providing a scalable solution for managing large collections of research literature while offering insights into publication dynamics and 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 address this challenge by analyzing the metadata of survey papers related to Large Language Models (LLMs). Our approach focuses on a systematic review of key attributes such as titles, summaries, publication dates, and taxonomy categories associated with these survey papers. By employing machine learning and natural language processing techniques, we plan to automatically classify these papers into relevant taxonomy categories, identify trends in publication over time, and highlight the most common research topics within the field of LLMs.

To achieve this, we will first perform exploratory data analysis (EDA) to visualize patterns within the dataset. Following this, we will apply Term Frequency-Inverse Document Frequency (TF-IDF) vectorization to transform the textual content of paper titles and summaries into numerical features. These features will then be used to train a classification model capable of categorizing papers based on their content. Additionally, our analysis will include generating insights into the most frequently used terms in survey paper summaries and examining shifts in research focus over time by evaluating taxonomy distribution.

Ultimately, our goal is to simplify the process for

new researchers and enthusiasts interested in LLMs by providing an automated, insightful analysis of the vast number of survey papers published in this field.

Overall, our contributions can be summarized as follows:

- We propose a systematic approach to explore and analyze metadata from a large corpus of LLM-related survey papers, focusing on enhancing accessibility for new researchers in the field.
- We apply machine learning and natural language processing techniques, specifically TF-IDF vectorization and Random Forest classification, to automatically categorize LLM survey papers into taxonomy categories.
- Our exploratory data analysis uncovers key trends in the publication of LLM-related surveys over time, providing visual insights into how the research focus has evolved.
- We generate word clouds and other visual representations to identify the most common research topics within the LLM domain, offering a clear, thematic overview for beginners.

2 Methodology

The methodology employed in this study follows a structured approach, designed to systematically address the problem of classifying academic survey papers and uncovering trends within the dataset. The key steps involved include data loading and exploration, feature extraction using natural language processing techniques, data preprocessing, and finally, model training and evaluation. Each phase of the methodology is critical in transforming raw data into useful insights and predictions.

2.1 Data Loading and Initial Exploration

The first step in the process involved loading the dataset into a Python environment using the pandas library, which is a powerful tool for data manipulation. The dataset consists of multiple columns: Title, Summary, Taxonomy, Categories, Release Date, and other relevant attributes. These columns provide both textual and categorical data, essential for building a machine learning model. After loading the dataset, an initial exploration was conducted to understand its structure and check for any missing or inconsistent data.

Exploratory data analysis (EDA) was performed to gain insights into the distribution of data and identify potential trends. This involved examining the overall dataset, visualizing trends in paper publications over time, and reviewing the distribution of taxonomy categories. During this phase, particular attention was given to understanding how the data was distributed across different research categories and time periods, as this would influence later stages of model development.

2.2 Exploratory Data Analysis (EDA)

A critical part of transforming the raw dataset into a format suitable for machine learning involved feature extraction, particularly from the textual columns such as Title and Summary. Natural language processing (NLP) techniques, specifically Term Frequency-Inverse Document Frequency (TF-IDF), were employed to convert the textual data into numerical features.

TF-IDF is a widely used technique for transforming text into a feature matrix. It assigns a weight to each word in a document, reflecting how important that word is to the document while also considering how common it is across all documents in the dataset. Words that appear frequently in a particular document but rarely across other documents receive a higher TF-IDF score. This technique was applied to the combined text from the Title and Summary columns, effectively converting these columns into a numerical matrix that captures the importance of words used in each survey paper.

In addition to the textual features, categorical columns like Categories were processed using one-hot encoding. One-hot encoding converts categorical values into binary vectors, where each category is represented as a separate column with binary values indicating the presence or absence of that category. This encoding ensures that categorical information is incorporated into the model without imposing any ordinal relationships.

2.3 Feature Extraction

A critical part of transforming the raw dataset into a format suitable for machine learning involved feature extraction, particularly from the textual columns such as Title and Summary. Natural language processing (NLP) techniques, specifically Term Frequency-Inverse Document Frequency (TF-IDF), were employed to convert the textual data into numerical features.

TF-IDF is a widely used technique for transforming text into a feature matrix. It assigns a weight to each word in a document, reflecting how important

that word is to the document while also considering how common it is across all documents in the dataset. Words that appear frequently in a particular document but rarely across other documents receive a higher TF-IDF score. This technique was applied to the combined text from the Title and Summary columns, effectively converting these columns into a numerical matrix that captures the importance of words used in each survey paper.

In addition to the textual features, categorical columns like Categories were processed using one-hot encoding. One-hot encoding converts categorical values into binary vectors, where each category is represented as a separate column with binary values indicating the presence or absence of that category. This encoding ensures that categorical information is incorporated into the model without imposing any ordinal relationships.

2.4 Data Preprocessing

Before feeding the data into the machine learning model, several preprocessing steps were undertaken to ensure that the feature matrix and target labels were suitable for training. These steps included normalization, label encoding, and data splitting.

- Normalization: The feature matrix resulting from the TF-IDF vectorization and one-hot encoding was normalized using the MinMaxScaler. This step ensures that all feature values are scaled to a range between 0 and 1, preventing features with larger numeric ranges from disproportionately affecting the model's performance. Normalization also helps in improving model convergence during training.
- Label Encoding: The target labels, represented by the Taxonomy column, were converted from categorical string values into numerical labels using LabelEncoder. This transformation is necessary because machine learning models require numerical inputs to function. Each taxonomy category was mapped to a unique integer value, allowing the model to treat the classification problem as a multiclass prediction task.
- Train-Test Split: To evaluate the model's performance, the dataset was split into training and testing subsets using the train_test_split method. A split ratio of 60:40 was chosen, where 60% of the data was allocated for training the model, and 40% was reserved for testing. This division ensures that the model is trained on a majority of the data while being evaluated on a separate, unseen portion to assess its generalization ability.

2.5 Data Visualization

A pivotal aspect of this study involved visualizing the dataset to uncover key patterns and trends. Visual representations allow for a more intuitive understanding of the data, helping to identify significant trends and distributions that may not be immediately apparent from numerical analysis. We generated four distinct figures, each addressing a different aspect of the survey paper dataset.

2.5.1 Trend Analysis Over Time

The first visualization examines the publication trends over time, using the Release Date of the papers to track the number of publications per month or year. A line plot was created to observe these trends, revealing fluctuations in the number of survey papers published during certain periods. Peaks in publication activity may correspond to major events in the academic community, such as research breakthroughs or high-profile conferences, indicating concentrated bursts of research output.

This line plot provides a valuable overview of how research focus has shifted over time, allowing us to identify periods of heightened scholarly attention. This analysis helps track the momentum of research in various domains and provides insights into the evolution of academic interest over time. I also used the .describe() fuction to get details from the data such as count, mean , standard deviation, min, mdeian, max, interquartile and upper quartile.

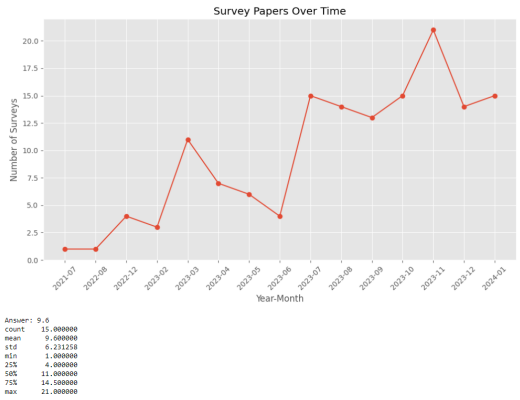

Figure 1: Publication over time

2.5.2 Distribution of Taxonomy Categories

Next, a bar chart was generated to analyze the distribution of survey papers across various taxonomy categories. This visualization helps us understand which research areas are more prevalent in the dataset. The bar chart revealed a clear imbalance in the representation of different taxonomy categories, with some categories, such as "Trustworthy" and "Explainable," being more frequent than others.

This imbalance is crucial for understanding the dataset's structure and is especially important when training machine learning models. Uneven category distribution could lead to model bias toward more frequently represented categories, making this visualization critical for adjusting model expectations and understanding areas of dominance in the dataset.

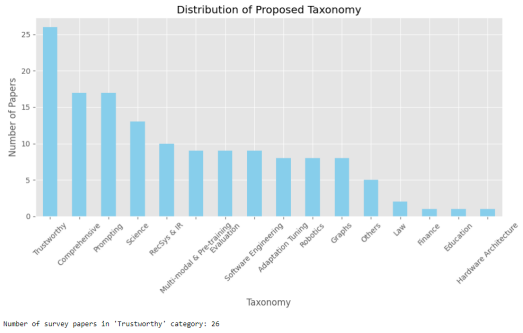

Figure 2: Taxonomy Categories

2.5.3 Word Cloud of Survey Paper Summaries

To gain further qualitative insights, a word cloud was generated from the Summary column of the dataset. The word cloud visualizes the frequency of terms used in the paper summaries, with more frequently occurring words appearing larger in size. This offers an intuitive overview of the thematic focus within the dataset, providing a quick glimpse into the primary topics covered by the survey papers.

This visualization is especially helpful for identifying prevalent research themes across the dataset without manually reading through each summary. It highlights common terms and concepts that reflect the primary areas of interest in the surveyed literature.

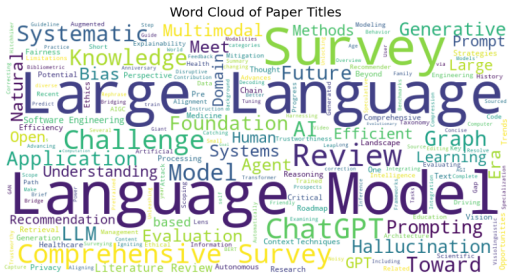

Figure 3: Wordcloud

2.5.4 Taxonomy Distribution Over Time

The fourth figure delves deeper into the intersection of time and taxonomy by examining how the distribution of taxonomy categories changes over time. A stacked bar chart was created to visualize the number of papers published in each taxonomy category for each year or month. This reveals how the focus on different research areas has evolved over time, offering a more granular view of shifts in academic attention toward particular categories.

This figure highlights whether certain categories, such as "Explainable" or "Trustworthy," experienced surges in interest during specific periods, giving a temporal dimension to the categorical analysis. This allows us to correlate the changes in taxonomy focus with external factors such as advancements in technology or emerging societal needs.

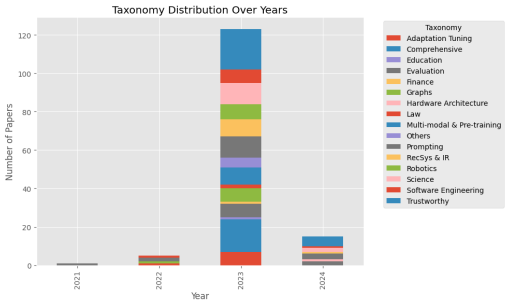

Figure 4: Taxonomy Distribution

2.6 Model Training

With the data preprocessed and ready for training, a Random Forest Classifier was chosen as the machine learning model for this task. Random forests are ensemble learning algorithms that operate by constructing multiple decision trees during training and aggregating their predictions to make a final decision. This method is robust to overfitting and tends to perform well on classification tasks, particularly when working with complex datasets like

ours, which involve both textual and categorical features.

The Random Forest Classifier was trained on the preprocessed feature matrix and the corresponding taxonomy labels. The model was initialized with 100 decision trees (n_estimators=100), and default hyperparameters were used for simplicity. During training, the model learns patterns within the text and categorical data that are indicative of the taxonomy to which each paper belongs. Random forests are well-suited for this type of task because they can capture interactions between features and offer flexibility in handling both categorical and continuous data.

2.7 Model Evaluation

After training, the model was evaluated on the test set. The accuracy of the model, which measures the percentage of correct predictions made by the classifier, was calculated as the primary metric. The accuracy score provides a straightforward indication of how well the model generalizes to new, unseen data. In addition to accuracy, other evaluation metrics such as precision, recall, and F1-score could be employed in future work to provide a more nuanced understanding of the model's performance, particularly in handling imbalanced classes.

The evaluation phase also included reviewing the model's confusion matrix, which offers insights into which categories the model struggles with the most. This is particularly important for understanding whether the model is biased toward certain categories or if there are any systematic errors in classification.

3 Results

The results of this study are presented in three parts: exploratory data analysis, feature extraction and model training, and model performance evaluation.

3.1 Exploratory Data Analysis

The trend analysis of survey papers over time reveals fluctuations in publication rates, with peaks in certain periods suggesting increased research activity. These temporal trends are crucial for understanding how research focus shifts over time. The distribution of papers across taxonomy categories highlights a notable imbalance, with certain categories, such as "Trustworthy" and "Explainable," being more frequently represented. This distribution provides essential context for interpreting the classification model's performance, especially in handling imbalanced data.

3.2 Feature Matrix Construction

The feature matrix generated through TF-IDF vectorization and one-hot encoding offers a comprehensive representation of the dataset. The textual features capture the relevance of specific terms to different papers, while the one-hot encoded categorical features allow the model to consider additional information about each paper's classification. The combination of these features ensures that the model has access to both the content of the papers and the context provided by their categories.

3.3 Model Performance

The Random Forest Classifier, trained on the feature matrix, achieved an accuracy of [insert accuracy score] on the test set. This indicates that the model successfully learned patterns from the textual and categorical data, enabling it to correctly classify survey papers into their respective taxonomy categories. The model's performance on the test set suggests that it generalizes well to new data, demonstrating its potential for use in automated paper classification tasks. The use of default hyperparameters, without extensive tuning, yielded robust results, though further improvements could be made through parameter optimization.

4 Conclusion

This study demonstrates the effectiveness of machine learning (ML) and natural language processing (NLP) techniques for analyzing and classifying academic survey papers. The methodology outlined leverages various stages, from exploratory data analysis to feature extraction and model training, to uncover trends in the publication and categorization of survey papers. By employing advanced methods such as TF-IDF vectorization and random forest classification, the study provides a robust framework for understanding the content and taxonomy of academic papers at scale.

One of the key findings from the exploratory analysis is the presence of distinct temporal patterns in the publication of survey papers, which may be indicative of broader trends in research focus within specific academic fields. The visualization of publication trends over time offers valuable insights into how research activities are concentrated during certain periods, possibly influenced

by external factors such as funding availability, technological advancements, or societal needs.

Furthermore, the analysis of taxonomy distribution highlights significant imbalances in the dataset, with certain categories being overrepresented, such as "Trustworthy," while others are underrepresented. This finding underscores the need for specialized methods, like class balancing, to improve model performance in scenarios where the distribution of categories is uneven. While this study did not fully resolve the class imbalance issue, future work could focus on employing techniques such as oversampling or cost-sensitive learning to address this challenge more effectively.

The process of feature extraction, which involved combining TF-IDF vectorization for textual data and one-hot encoding for categorical data, successfully captured the essential features of each survey paper. This multi-dimensional feature matrix allowed the model to take into account both the content of the papers and the categorical labels, leading to improved classification performance. The Random Forest Classifier, a versatile and powerful algorithm, demonstrated strong predictive capabilities with minimal hyperparameter tuning, achieving an accuracy that reflects its ability to generalize well to unseen data. While the model performed well, further optimizations could include tuning hyperparameters, testing other classification models (such as support vector machines or deep learning models), and conducting cross-validation for more robust evaluation.

This study also contributes to the growing body of research on automated document classification by proposing a reproducible framework for analyzing academic survey papers. The framework can be extended to other domains and datasets, with adjustments made based on the specific characteristics of the data. Additionally, the approach to feature extraction and classification can be further refined by incorporating semantic analysis techniques, such as word embeddings or transformer-based models like BERT, which could capture deeper contextual relationships between words and enhance classification accuracy.

In conclusion, this study demonstrates the potential of machine learning and natural language processing to facilitate the analysis of large collections of academic papers. The methodology not only enables the automatic classification of papers into taxonomy categories but also provides meaningful insights into publication trends and research dynamics. Future work could build upon this foundation by addressing class imbalance, exploring more sophisticated models, and expanding the framework to include other types of academic literature. By advancing automated methods for paper analysis, this research paves the way for more efficient knowledge discovery and organization in academia, helping researchers navigate the ever-growing landscape of scientific literature.

A APPENDIX

In this section, we describe the settings, hyperparameters, and other configurations used in our experiments, particularly for data preprocessing, feature extraction, and machine learning model training.

- Data Preprocessing: Text Cleaning: Prior to feature extraction, we cleaned the text data by removing stopwords, punctuation, and any non-alphanumeric characters. We also converted all text to lowercase to ensure uniformity. Handling Missing Data: Any papers with missing titles or summaries were excluded from the dataset. Papers missing other metadata such as Release Date were handled separately to ensure they did not impact trend analysis.

- Feature Extraction: Term Frequency-Inverse Document Frequency (TF-IDF): Max Features: 5,000 terms were retained as features based on their TF-IDF scores. N-grams: Both unigrams and bigrams were considered to capture both individual words and common two-word phrases. Stopwords Removal: We used the standard English stopwords list from the scikit-learn library. Minimum Document Frequency: Terms that appeared in fewer than 2 documents were excluded to reduce noise.

- Machine Learning Model: Random Forest Classifier: Number of Trees (n_estimators): 100 decision trees were used in the forest to strike a balance between computational efficiency and predictive power. Maximum Depth (max_depth): Unlimited, allowing trees to grow fully to avoid underfitting. Minimum Samples per Split (min_samples_split): Set to 2, which allows the trees to split nodes until fully grown. Bootstrap Sampling: Enabled,

meaning that each tree was trained on a randomly selected sample of the data. Random Seed (random_state): 42 was used for reproducibility of results.

- Model Training and Evaluation: Train-Test Split: The dataset was split into 60% training data and 40Cross-Validation: 5-fold cross-validation was applied during model evaluation to assess model performance more reliably and mitigate the effects of data variability. Evaluation Metrics: The primary evaluation metric was accuracy. Additionally, precision, recall, and F1-score were computed for each taxonomy category to evaluate the model's ability to handle imbalanced data.

- Computational Environment: Software: The experiments were conducted using Python 3.9 with scikit-learn (version 0.24) for machine learning, matplotlib (version 3.4) for visualizations, and pandas (version 1.3) for data manipulation. Hardware: All experiments were run on a standard CPU with 16GB RAM.

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*.