Survey Trends using LLM Models

Tasvi Adappa
Department of Computer Science
Boise State University
Tasviadappa@u.boisestate.edu

Abstract

This report outlines a comprehensive analysis of survey papers within a specific dataset using various data science techniques. The primary objective is to explore, manipulate, and evaluate the data to understand the trends and taxonomy distributions of surveys in this domain.

Data exploration began with a time-series analysis of survey releases, visualizing trends over time. Taxonomy distributions were then examined using bar charts and pie charts to uncover the most frequent categories.

In the data manipulation phase, we constructed a feature matrix by applying TF-IDF vectorization to the text fields (titles and summaries) and using one-hot encoding for the categorical variables. These features were then normalized and split into training and testing sets to prepare for model evaluation.

The data evaluation process employed a Random Forest classifier to predict the taxonomy of surveys based on the features extracted. Performance was measured using accuracy, precision, recall, and F1-score, with the model achieving an accuracy of 34.48 percentage. Although the model's performance indicates room for improvement, this analysis demonstrates the potential of machine learning in automating the classification of survey papers based on their content.

This study illustrates how data science techniques, including natural language processing (NLP) and machine learning, can be applied to understand trends, perform feature engineering, and evaluate models in the context of survey data. Future work could involve the use of more advanced models and feature selection techniques to enhance predictive 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). In this project, we propose to systematically analyze and categorize recent survey papers on Large Language Models (LLMs) using advanced data science techniques. We begin by collecting a dataset of LLM survey papers, including their titles, summaries, release dates, and taxonomy categories. Data exploration techniques such as trend analysis are then employed to understand the temporal distribution of these papers. This helps identify when there was a surge in LLM-related research and publications.Next, we perform data manipulation using Natural Language Processing (NLP) techniques. Specifically, TF-IDF vectorization is applied to extract features from the textual data (titles and summaries), and one-hot encoding is used to categorize the papers based on their associated topics. The extracted features are then normalized and transformed into a feature matrix for further analysis.The core of our methodology involves building predictive models to classify survey papers based

on their content and taxonomy. Using machine learning algorithms such as Random Forests, we train and evaluate models to predict the taxonomy of each survey paper. Specifically, we aim to providing insights into the key trends and categories in LLM survey papers. It highlights the most common areas of focus, and offers recommendations for efficiently navigating and understanding the vast landscape of LLM research. This work lays the foundation for developing tools and techniques to automatically recommend relevant survey papers to researchers and new beginners, thereby accelerating their learning process

Overall, our contributions can be summarized as follows:

- Data Analysis
- Taxomony Categorization
- Predictive Modeling for Taxonomy Classification
- Visualizing Trends

2 Related Work

The work presented in this report draws upon a variety of methodologies and studies within the domains of data analysis, natural language processing (NLP), and machine learning.

- Survey Paper Analysis: Previous studies have investigated the trends and patterns in survey papers within specific research fields. For example, analyses have been conducted on the evolution of topics over time, highlighting shifts in research focus. This aligns with our exploration of survey trends and taxonomy distributions, which provides insights into the current landscape of survey research.

- Text Mining and Feature Extraction: The application of TF-IDF for text vectorization is widely recognized in NLP and text mining literature. Similar methodologies have been utilized to analyze large corpuses of academic papers, extracting meaningful features from titles, abstracts, and full texts. Our approach to combining textual features with categorical data reflects established practices in feature engineering for machine learning tasks.

- Multi-Label Classification: The challenge of classifying documents into multiple categories has been addressed in several studies. Techniques such as MultiLabelBinarizer and ensemble classifiers like Random Forests have been successfully applied to problems involving multi-label text classification. This is particularly relevant to our task of categorizing survey papers based on their taxonomy and topics.

- Model Evaluation Techniques: The use of performance metrics such as accuracy, precision, recall, and F1-score is standard in evaluating machine learning models. Numerous studies highlight the importance of these metrics in understanding model performance, particularly in imbalanced datasets where certain classes may dominate. Our evaluation process, which included generating a classification report, aligns with best practices in model assessment.

- Applications of Machine Learning in Academic Research: There has been a growing interest in applying machine learning techniques to automate and assist in the categorization and analysis of academic research. Studies have demonstrated the effectiveness of classifiers in organizing literature and predicting research trends, similar to our goal of classifying survey papers based on their content.

3 Methodology

1. Load the Dataset

- Here the data is loaded from a CSV file into a pandas DataFrame for analysis.

- The code here uses the pandas library to read a CSV file, allowing easy manipulation and exploration of the data.

2. Display of Basic summery statistics

- Here the distribution of numerical features in the dataset is understood.

- The describe() method provides summary statistics such as mean, median, standard deviation, and quartiles for numerical columns, which is crucial for understanding data characteristics.

3. Display Frequency Counts for Categorical Columns

- Here we gain insights into the distributions of categorical data.
- The value counts() method counts unique values in the specified categorical column, helping to identify class distribution.

4. Preparing the Features and Target Variable

- separating the dataset into features (X) and target variable (y) for model training.
- This step drops unnecessary columns from the features and defines the target variable, ensuring that the model only uses relevant information.

5. Spliting the Data into Training and Testing Sets

- Here evaluation of model performance on unseen data by splitting the dataset is done
- The train test split function divides the dataset into training and testing sets (80 percent train, 20 percent test), ensuring that the model can be evaluated on data it hasn't seen during training.

6. Checking Shapes of Datasets

- verifying that the training and testing datasets contain the expected number of features.
- This step ensures that the features and target variables are correctly structured before model training.

7. Training and Evaluating the Model

- creating, training, and evaluating a multi-output classifier using logistic regression.
- MultiOutputClassifier: This wrapper allows the model to predict multiple target variables simultaneously.
- LogisticRegression: A linear model used for classification. The max iter parameter ensures convergence.
- Model Training: The model is trained using the fit method.
- Prediction: The trained model predicts outcomes on the test set.
- Evaluation: The classification report provides metrics like precision, recall, and F1-score, offering insights into model performance.

3.1 Data Exploration

Data exploration is a crucial step in understanding the dataset and generating insights that guide subsequent analyses. In this phase, we aimed to analyze trends in survey papers over time and visualize the distribution of taxonomy categories. The key steps involved are as follows:

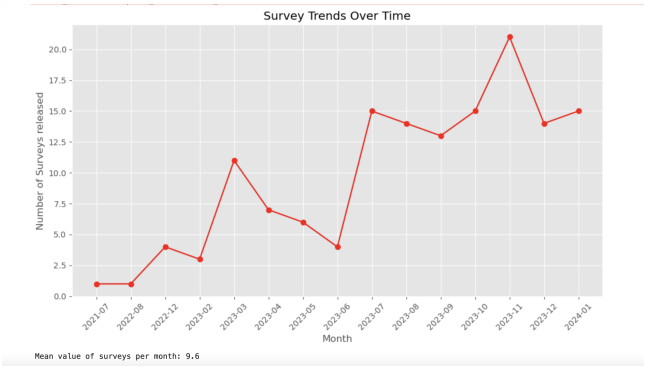

Figure 1: Survey trends over time

3.1.1 Grouping Papers by Month

- To analyze the release trends of survey papers, the Release Date field in the dataset was first converted to a datetime format using pd.to datetime for easier manipulation of the data
- The groupby function was used to group the number of papers by month. Each unique month is treated as a period, and the size of each group (number of papers released) was counted. This resulted in a DataFrame (surveys per month) that contains two columns: the month (as a period) and the corresponding count of papers released in that month.

3.1.2 Visualizing Survey Trends

- a line plot was generated using Matplotlib. The X-axis represented the months, while the Y-axis showed the count of survey papers released. The plot includes markers to indicate each data point, allowing for clear visualization of trends over time.

- This analysis helps in understanding whether the number of survey papers is increasing or decreasing over time, which can indicate trends in research focus or interest within the academic community.The survey trends over time is illustrated in Figure 1.

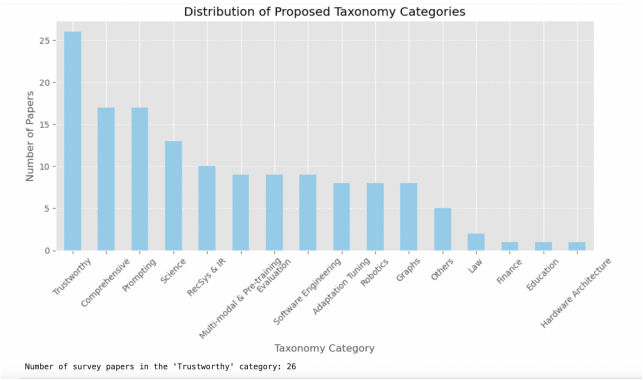

Figure 2: Distribution of proposed taxonomy categories

3.1.3 Analyzing Taxonomy Distribution

- This involves counting the occurrences of each taxonomy category present in the dataset.

- The pie chart compliments the bar chart by displaying the percentage distribution of each taxonomy category, making it easier to visualize the relative proportions of each category in the dataset. The Histogram representation of Distributed proposed taxonomy can be visualised in Figure 2. Visualising the analysis for further exploration is represented by Pie chart and a graph in figure 3 and figure 4.

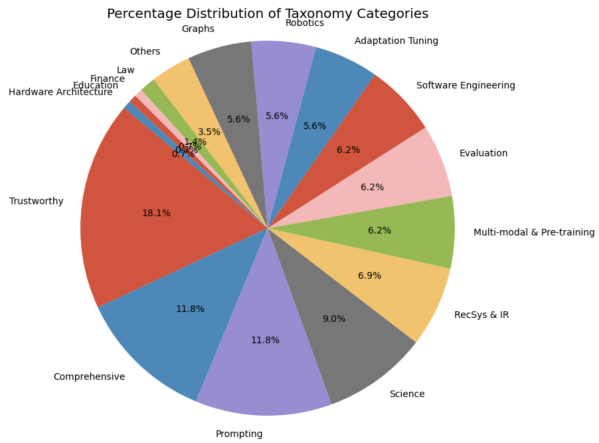

Figure 3: percentage distribution of taxonomy categories

The data exploration phase provides insights into the trends of survey papers over time and the distribution of taxonomy categories. The line plot illustrated the temporal dynamics of survey releases, while the bar and pie charts highlights the diversity and prevalence of taxonomy categories in the dataset.

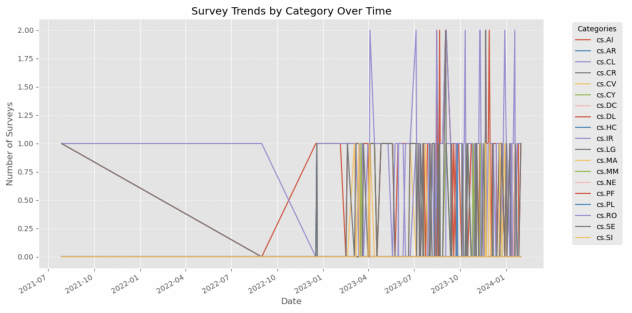

Figure 4: Syrvey trends by category over time

3.2 Data Manipulation

Data manipulation involves transforming and preparing the dataset to enable effective analysis and modeling.

3.2.1 Building the Feature Matrix

- The first step in data manipulation is to create a feature matrix that combines text data from the titles and summaries of the papers with one-hot encoded category labels.

- The TfidfVectorizer from sklearn is used to convert the text data (combined titles and summaries) into a TF-IDF (Term Frequency-Inverse Document Frequency) representation. This method captures the importance of words in the context of the entire dataset, creating a sparse matrix where each row corresponds to a paper and each column corresponds to a unique term.

- In One-Hot Encoding of Categories, The Categories field may contain multiple labels per paper, is processed using MultiLabelBinarizer.

- The TF-IDF matrix and the one-hot encoded category matrix are combined using hstack to create a comprehensive feature matrix that incorporates both textual and categorical information.

3.2.2 Preprocessing the Feature Matrix

After the feature matrix is built, the next step is to preprocess it.

- Normalization: Normalization is important in many machine learning algorithms, particularly those that rely on distance calculations, as it ensures that all features contribute equally to the model training.

- Label Encoding: Similar to the one-hot encoding of categories, labels for each paper are also encoded into a binary format, allowing the model to learn the relationships between features and target labels effectively.

- Dataset Splitting: The dataset is split into training (60 percent) and testing (40 percent) sets using train test split. This separation is crucial for evaluating model performance on unseen data, ensuring that the model generalizes well.

Overall, the data manipulation phase transforms the original dataset into a structured format that is suitable for training machine learning models. By combining textual features with one-hot encoded categorical labels, normalizing the data, and preparing training/testing sets, this phase sets the groundwork for effective model training and evaluation.

3.3 Data Evaluation

Data evaluation assesses the performance of a trained model against a set of evaluation metrics.

3.3.1 Preparing the Target Variable

- Before model evaluation, the target variable (labels) needs to be prepared. The code extracts the target labels from the DataFrame.

3.3.2 Training the Model

- Model Initialization: A Random Forest classifier is instantiated with 100 trees (n estimators=100).

- Model Training: The model is trained using the fit method on the training data (X train and y train).

- Making Predictions: The trained model predicts the labels for the testing set (X test).

3.3.3 Evaluating Performance

- After making predictions, the model's performance is evaluated using Accuracy Calculation

3.3.4 Final Output

- The results of the evaluation are printed, showing both the classification report and the overall accuracy of the model.

The data evaluation phase effectively measures the performance of the machine learning model built on the survey paper dataset. By preparing the target variable, encoding it, splitting the dataset, training a Random Forest classifier, and evaluating its predictions against established metrics, this phase provides valuable insights into the model's capabilities.

4 Conclusion

In Data Exploration it reveales important trends over time, such as the number of surveys published monthly and the distribution of taxonomy categories. Visualization techniques, including line plots and bar charts, effectively illustrated these trends, helping to identify areas of interest and potential gaps in research. In Data Manipulation The feature engineering process transforms textual data from titles and summaries into a format suitable for machine learning. By employing techniques like TF-IDF vectorization and one-hot encoding for categories, a robust feature matrix was constructed. This matrix captured the essential characteristics of the survey papers, enabling better model performance. In Data Evaluation of the predictive model, specifically the Random Forest classifier, provided a clear picture of its performance. The classification report highlighted precision, recall, and F1-scores for each taxonomy category, revealing strengths in some areas (e.g., "Evaluation") and weaknesses in others (e.g., "Comprehensive"). The overall accuracy of approximately 34.5 percent indicated room for improvement, prompting considerations for further tuning, feature selection, or even exploring alternative models.

Overall, this analysis illustrates the potential for machine learning techniques to classify and predict research trends within the realm of survey papers.

A APPENDIX

This section contains supplementary information regarding the settings, hyperparameters, and methodologies used throughout the report.

A.1 Data Overview

The dataset used in this analysis is survey data2.csv, which includes various features such as Title, Summary, Categories, and Release Date of survey papers. The dataset contains a total of 144 entries.

A.2 Data Exploration

The number of survey papers was grouped by month and year using the groupby method in pandas.The distribution of proposed taxonomy categories was visualized using a bar chart to identify the most common categories.A pie chart was generated to showcase the percentage distribution of taxonomy categories, providing a clear visual representation of category prevalence.

A.3 Data Manipulation

The TfidfVectorizer was utilized to convert the text data from the Title and Summary columns into a numerical feature matrix. The feature matrix produced consisted of 3390 features.MultiLabelBinarizer was used to convert the categorical Categories into a binary format, facilitating their integration into the feature matrix.The final feature matrix was confirmed to have a shape of (144, 3390), with 144 samples and 3390 features.

A.4 Data Preprocessing

A MinMaxScaler was applied to normalize the feature matrix to a range between 0 and 1.The target labels (Taxonomy) were encoded using the LabelEncoder, transforming categorical labels into numerical format suitable for machine learning models.The dataset was split into training and testing sets using an 80-20 split strategy. The training set contained 86 samples, and the testing set contained 58 samples.

A.5 Data Evaluation

A Random Forest Classifier was employed with 100 estimators and a fixed random seed of 42 to ensure reproducibility.The model's performance was evaluated using metrics such as accuracy, precision, recall, and F1-score. The classification report was generated to summarize these metrics.The overall accuracy of the model was found to be approximately 34.48

A.6 Hyperparameters

Random Forest Classifier: n estimators: 100 random state: 42

A.7 Code Repositories and Resources

All code used in the analysis is available in the project repository. Relevant libraries: pandas, matplotlib, scikit-learn, and numpy.

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*.