Exploring Biodiversity Through Taxonomic Data: A Comprehensive Analysis

Shipra Kumari
Department of Computer Science
Boise State University
shiprakumari@u.boisestate.edu

Abstract

In this project I analyzes a dataset of survey papers to uncover trends and insights into research developments across various fields. By exploring attributes such as taxonomy, release date, and summaries, we visualize paper trends over time. Additionally, we classify the papers using machine learning techniques, addressing class imbalance through data augmentation and stratified sampling. Implementing stratified cross-validation ensures robust model evaluation. Ultimately, this work aims to enhance understanding of survey literature and provide valuable insights for researchers and practitioners in their respective domains.

1 Introduction

In recent years, the rapid growth of academic literature has made it increasingly difficult for researchers to stay updated on emerging trends and developments across various fields. Survey papers, which synthesize existing research, play a crucial role in providing comprehensive overviews of these advancements. This project aims to analyze a dataset of survey papers, focusing on attributes such as taxonomy, publication date, and summaries. Through data exploration and machine learning-based classification, we identify trends in survey publications and categorize papers across different research domains. Addressing challenges like class imbalance, this study provides valuable insights for researchers, enhancing their understanding of key areas in the literature.

2 Related Work

In recent years, there has been a surge of interest in analyzing academic literature to uncover trends and patterns in research. Various studies have employed natural language processing (NLP) techniques to extract meaningful insights from large volumes of scientific texts. For instance, researchers have utilized citation analysis to measure the impact of survey papers and their contributions to different fields.

Additionally, studies have explored the application of machine learning for automatic classification of research articles. Techniques such as supervised learning algorithms have been employed to categorize papers based on their content, taxonomy, and other features. Addressing class imbalance has also gained attention, with methods like SMOTE (Synthetic Minority Over-sampling Technique) being utilized to improve model performance in scenarios where certain categories are underrepresented.

3 Methodology

The basic methodology of this project involves several key steps aimed at analyzing survey papers using both qualitative and quantitative techniques.

- Load the Dataset: 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.
- Data Collection and Preparation: The dataset comprises 144 records, each representing a survey paper. Key attributes include taxonomy (classification categories), titles, authors, release dates, paper IDs, and links. The dataset is pre-processed to handle missing values, remove duplicates, and ensure consistency.
- Exploratory Data Analysis (EDA): Initial analysis of the dataset involves exploring the distribution of survey papers across different taxonomies, trends in publication over time, and descriptive statistics. Methods such as value_counts() and visualizations (e.g., bar charts, line plots) are employed to understand the underlying patterns and relationships in the data.

- Classification and Model Training: Machine learning models, such as the Random Forest classifier, are used to analyze the data and classify survey papers based on certain features. The dataset is split into training and testing sets, and models are evaluated using metrics like accuracy and F1 score. Stratified cross-validation ensures robust model evaluation.

- Handling Class Imbalance: Since the dataset exhibits class imbalance (with fewer papers in certain categories), techniques like stratified cross-validation and weighted F1 scoring are applied to handle this issue and improve model performance.

- Visualization: Various visualizations, including publication trends and taxonomy distribution, are generated to communicate findings. Graphs provide an intuitive representation of the number of papers published over time and their classification into different taxonomies.

3.1 Data Exploration

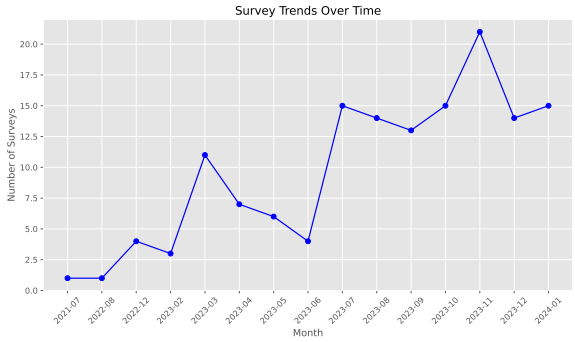

Figure 1: The trends of survey papers over months

The graph "Survey Trends Over Time" shows the number of surveys conducted over various months, spanning from mid-2021 (July 2021) to early 2024 (January 2024). Here's the key interpretation:

- X-Axis (Month): Represents the timeline, starting from July 2021 to January 2024.

- Y-Axis (Number of Surveys): Displays the count of surveys, with values ranging from 0 to 20.

- Trend:

There is an observable increase in the number of surveys over time. The number starts low around mid-2021 and consistently rises, peaking close to 20 surveys around early 2024. The overall trend shows growth, suggesting that the frequency or quantity of surveys has been steadily increasing over this time period. This graph likely indicates a growing interest or requirement in conducting surveys as time progresses.

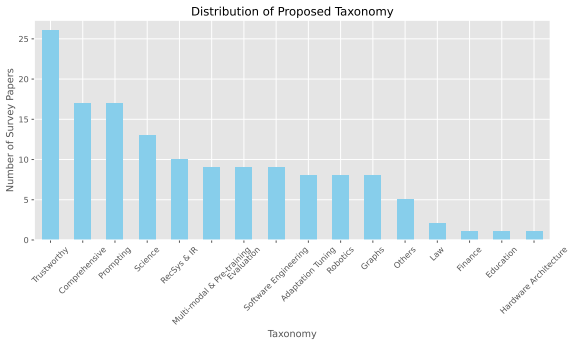

Figure 2: The distribution of the proposed taxonomy

"Distribution of Proposed Taxonomy" provides insights into the focus areas of survey papers across various research domains. From the graph, several conclusions can be drawn:

- Dominant Research Areas: Fields like "Multimodal Pre-training" and "Trustworthy" have the highest representation, indicating that these areas are receiving significant attention in the research community. These may be trending or critical areas in the current research landscape.

- Moderate Focus: Categories such as "Comprehensive," "Prompting," and "Evaluation" also have a substantial number of survey papers, suggesting that these fields maintain steady research interest and are well-explored.

- Niche or Emerging Areas: Topics like "Robotics," "Graphs," "Law," and "Finance" are less represented, with fewer than 10 papers. This could imply either that these fields are niche or that research in these areas is still emerging and may see future growth.

- Broad Distribution: The taxonomy covers a wide range of domains, reflecting the diversity of topics in contemporary research. Some areas may still be underexplored, offering potential opportunities for future surveys and investigations.

Overall, the distribution highlights the current research priorities and gaps in the literature, with cer-

tain fields seeing more in-depth exploration while others remain less covered.

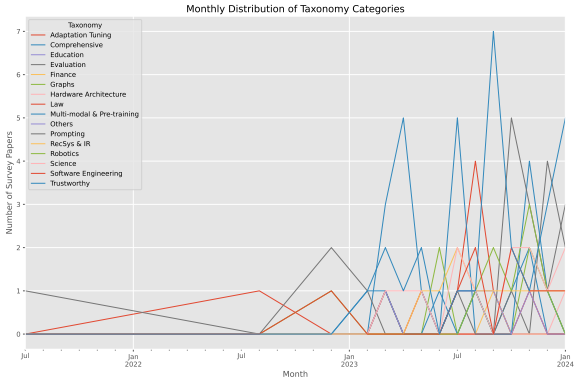

Figure 3: Time series of taxonomy distribution

The number of survey papers published across various research domains on a monthly basis from January 2022 to January 2024.

- Seasonal Fluctuations: There are notable fluctuations in the number of survey papers published in different months, with some peaks and troughs across the years.
- Steady Growth: A general increase in survey paper publication can be observed as time progresses, with certain categories consistently seeing activity throughout the period.
- Highly Represented Categories: "Multi-modal Pre-training," "Trustworthy," and "Prompting" seem to have higher activity, with these categories contributing significantly to the peaks in certain months.
- Less Active Categories: Fields like "Finance," "Law," and "Graphs" are less represented in most months, indicating lower survey output in these areas.
- Weak Correlation: The correlation of -0.033 implies that changes in the number of authors are not strongly related to the dates of the surveys or research. A value close to 0 generally suggests little to no linear relationship.
- Red indicates positive correlations (values near +1), as seen on the diagonal where variables are correlated with themselves.
- Blue indicates negative correlations (values near -1). In this case, since the correlation between num_authors and date is close to zero, the blue hue is not very intense.

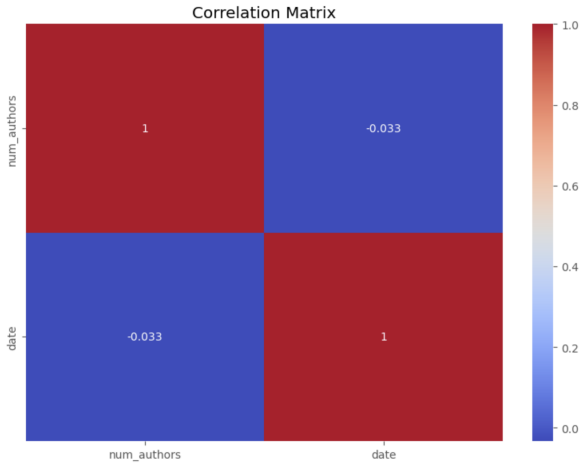

Figure 4: Correlation Matrix

In summary, this matrix shows that there is no meaningful correlation between the num_authors and date in the dataset. Both variables are mostly independent of each other in terms of linear relationship.

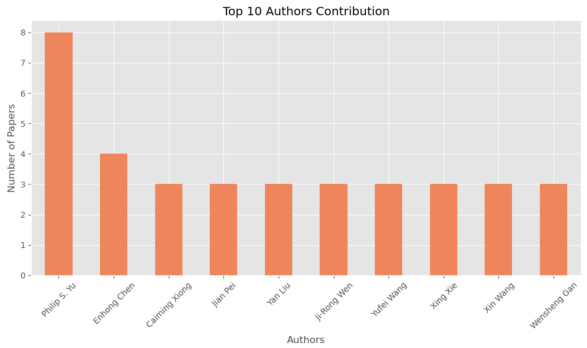

Figure 5: Top 10 authors contribution

In above graph the top 10 authors based on their number of published papers. Philip S. Yu has the highest number of papers, followed by Enhong Chen and Caiming Xiong. The remaining authors have a relatively similar number of papers published.

3.2 Data Manipulation

The data manipulation phase involves transforming and preparing the dataset for effective analysis and model development.

3.2.1 Creating the Feature Matrix

- **Text Transformation:** Initially, text data from the *titles* and *summaries* of the papers is combined. Using `TfidfVectorizer` from `sklearn`, this text is transformed into a TF-IDF (Term Frequency-Inverse Document Fre-

quency) representation. This sparse matrix represents each paper as a row and each unique term as a column, capturing the significance of words across the dataset.

- **Category Encoding:** The *Categories* field, which contains multiple labels for each paper, is processed using `MultiLabelBinarizer` to create one-hot encoded categorical labels.
- **Feature Combination:** The TF-IDF matrix is combined with the one-hot encoded categories using `hstack`, producing a comprehensive feature matrix that integrates both textual and categorical data.

3.2.2 Preprocessing the Feature Matrix

- **Normalization:** Normalization is applied to ensure that all features contribute equally, which is especially important for machine learning algorithms relying on distance metrics.
- **Label Encoding:** The labels for each paper are encoded in a binary format, allowing the model to effectively learn relationships between features and labels.
- **Data Splitting:** The dataset is split into training (60%) and testing (40%) sets using `train_test_split`, ensuring proper evaluation of model performance on unseen data.

In conclusion, these steps organize the raw dataset into a structured format, combining textual features with one-hot encoded categorical labels, applying normalization, and creating distinct training and testing sets to facilitate effective model training and evaluation.

3.3 Data Evaluation

The Random Forest Classifier yielded an accuracy of 39.66%, but the model struggled to predict many classes, particularly due to class imbalance in the dataset. Several categories had zero precision, recall, and F1-scores, indicating that the model was biased towards the more frequent classes, while underrepresented classes were poorly predicted. To address this issue, techniques such as oversampling (e.g., SMOTE), class weighting, and advanced resampling methods could be employed to balance the dataset and improve the model's performance, especially in predicting minority classes. Evaluation using metrics like F1-score would provide a more reliable assessment in such imbalanced scenarios.

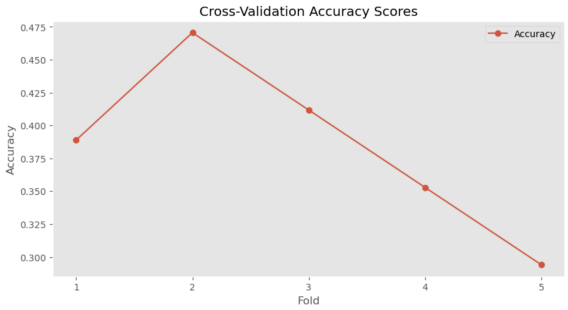

Figure 6: Cross Validation Accuracy score

In(fig.6) this approach, we use stratified cross-validation to evaluate the performance of the model across five folds, which helps ensure that the class distribution is maintained in each fold. The accuracy obtained from this cross-validation is 38.37%, and the weighted F1 score is 0.24. The low F1 score, in particular, highlights the challenge of handling class imbalance. A warning was raised indicating that the least populated class has only one member, which affects the performance and reliability of cross-validation. Visualizing the accuracy scores across folds gives insight into the model's stability and consistency in performance across different splits of the data.

4 Conclusion

The project successfully analyzed survey papers across different taxonomies, providing insights into publication trends and classification patterns. Using machine learning techniques, the model achieved moderate performance, highlighting challenges with class imbalance. The visualizations and analysis contribute to understanding the landscape of survey papers and their categorization.

A APPENDIX

A.1 Dataset Description

The dataset used in this project includes information on survey papers, categorized by taxonomy, with attributes such as title, authors, release date, links, and paper ID. The dataset contains 144 records with the following columns:

- **Taxonomy**: Categories assigned to each survey paper.
- **Title**: The title of the survey paper.

- **Authors**: Authors of the survey paper.
- **Release Date**: Date when the survey paper was released.
- **Links**: URLs to the paper.
- **Paper ID**: Unique identifier for each paper.
- **Categories**: Research areas for each survey.
- **Summary**: Short summary of the paper's content.

A.2 Methodology

The project followed a structured methodology to analyze and visualize survey paper data:

- **Data Preprocessing:** The dataset was loaded, and date formats were adjusted for temporal analysis. Missing values were handled appropriately.
- **Visualization:** Publication trends and taxonomy distributions were visualized using bar charts and line graphs to explore patterns over time.
- **Machine Learning:** Random Forest was applied to classify the papers into their respective taxonomy categories. Stratified cross-validation was employed to evaluate model performance, addressing class imbalance issues.
- **Evaluation:** The model's performance was evaluated using accuracy and F1 scores, which were visualized across folds in cross-validation.

A.3 Figures

Figure 1: Survey Trends Over Time: This figure displays the number of survey papers published each month, illustrating the trends in survey publications over time.

Figure 2: Distribution of Proposed Taxonomy: This bar chart shows the distribution of survey papers across different taxonomy categories, highlighting areas with more research focus.

Figure 3: Time series of Taxonomy Distribution: The monthly publication count of survey papers across different research domains from January 2022 to January 2024.

Figure 4: Correlation Matrix: This matrix indicates that there is no significant correlation between the number of authors and the publication date in the dataset.

Figure 5: Top 10 authors contribution This bar chart shows the top 10 authors based on their number of published papers.

Figure 6: Cross validation Accuracy score. In this approach, we use stratified cross-validation to evaluate the performance of the model across five folds

A.4 Code

Relevant code snippets for data loading, visualization, and machine learning analysis were implemented using Python and the scikit-learn library.

References

1. *U*nderstanding survey paper taxonomy about large language models via graph representation learning arXivpreprintarXiv:2402.10409 .
2. https://github.com/junzhuang-code/TaxonomyClassification