Data Exploration, Manipulation, and Evaluation

Sunny Moran Jr
Department of Computer Science
Boise State University
sunnymoran@u.boisestate.edu

Abstract

This paper explores various methods for analyzing, manipulating, and evaluating survey data. Using a Jupyter notebook, we examine trends such as the increase in the number of published survey articles over the years. Additionally, we categorize each article into different taxonomy topics. Through data manipulation, we create a TF-IDF matrix, enabling us to predict an article's taxonomy category based on key words from its title, summary, and associated categories

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, our goal is to provide visual models and examples to analyze and break down the information at hand. Using Jupyter notebooks, I will work with data from documents to create visualizations. This includes showing the average monthly publication of articles, generating TF-IDF matrices, and evaluating the data to assess the accuracy of each article's classification.

I believe my contributions demonstrate that, despite the complexity of large language models (LLMs), using Jupyter notebooks allows us to explore their relationship with machine learning models more effectively. This approach helps uncover trends in our data, such as predicting the taxonomy labels of documents based on the words in their titles, summaries, and categories.

2 Methodology

2.1 Data Exploration

In the initial steps of data exploration, we define functions to group published survey papers by month and use basic plotting functions to visualize trends in the data. Between 2021 and 2024, we observe a significant rise in the number of published survey papers on a monthly basis. By creating custom functions, we can process survey data from CSV files and generate visualizations that help predict future trends in paper publications. Additionally, our Jupyter notebook allows us to analyze quartile values (25-75) for the number of surveys published per month.

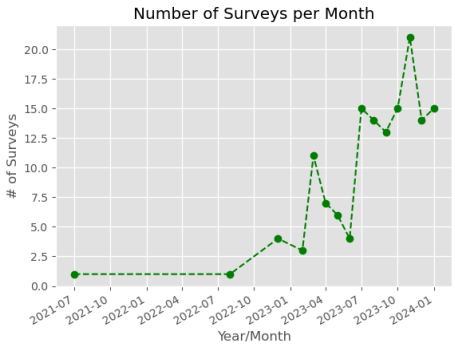

Figure 1: Number of surveys per month

Beyond identifying the growth in survey papers,

we analyze the distribution of the data to propose a taxonomy. Using bar charts, we visualize which categories have the highest number of survey papers between 2021 and 2024, with "trustworthy" emerging as the leading category. Another useful approach is aggregating data by taxonomy and publication date to get the total count of papers in each category. This method provides more detailed insights for users seeking exact numbers.

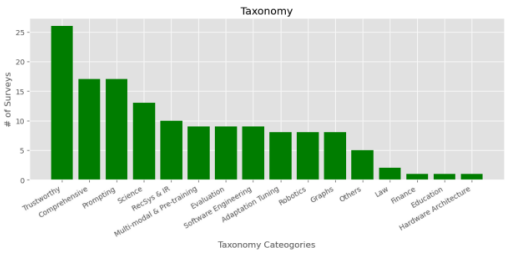

Figure 2: Number of surveys per month

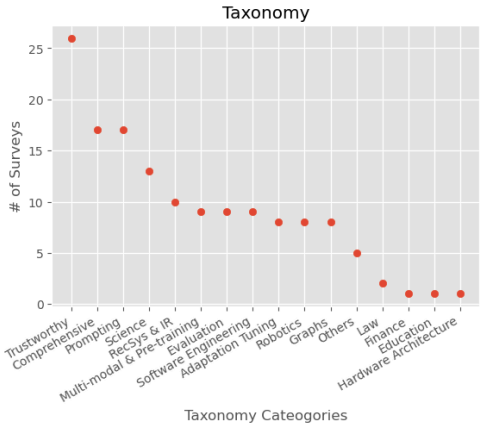

Figure 3: Number of surveys per month

Overall, data exploration offers a variety of visualization tools, such as bar charts and scatter plots, that help both scientists and end-users identify trends and patterns in the data, particularly in understanding which categories hold the most survey papers.

2.2 Data Manipulation

After completing the data exploration phase, we transition to creating a TF-IDF matrix, which associates TF-IDF values with corresponding words in the survey documents. This will help us evaluate and draw conclusions on the proposed taxonomy. A key challenge in this process was handling CSV data, where determining which values are separate and which are grouped together can be tricky—particularly in the "categories" section. For instance, some documents contain multiple categories like "cs.IR," "cs.LG," and "cs.MM," but commas between them are not treated as delimiters. This complicates methods like one-hot encoding, which rely on cleanly separated values for accurate category association.

	000	100	109	116	118	134	15	151	155	16	...	cs.IR	cs.LG	cs.MA
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	...	False	True	False
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	...	False	False	False
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	...	False	False	False
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	...	False	True	False
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	...	False	False	False
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
139	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	...	False	False	False
140	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.067178	...	False	False	False
141	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	...	False	True	False
142	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	...	False	False	False
143	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	...	False	False	False

Figure 4: Number of surveys per month

To resolve this, I used techniques such as splitting the category strings and applying a dataframe "explode" function to separate categories properly, ensuring more accurate results during data evaluation. After building the feature matrix, we normalize the data using a MinMaxScaler and encode the labels to incorporate the taxonomy. We then split the dataset with a 60-40 training-test ratio. With this step complete, we are prepared for further data manipulation.

2.3 Data Evaluation

There are several machine learning models that can be used to analyze our TF-IDF model, and for data evaluation, I chose the RandomForestClassifier. It is well-suited for classification tasks as it builds multiple decision trees on different data samples and improves accuracy by averaging the results.

```
Random Forest Accuracy: 34.48%
Classification Report for Random Forest:
              precision    recall  f1-score   support

           0       0.00      0.00      0.00         4
           1       0.25      0.12      0.17         8
           3       0.50      0.67      0.57         3
           5       0.00      0.00      0.00         4
           8       0.00      0.00      0.00         6
          10       0.43      0.50      0.46         6
          11       0.50      1.00      0.67         1
          12       0.00      0.00      0.00         4
          13       1.00      0.29      0.44         7
          14       1.00      0.25      0.40         4
          15       0.26      0.91      0.41        11

    accuracy                           0.34        58
   macro avg       0.36      0.34      0.28        58
weighted avg       0.35      0.34      0.27        58
```

Figure 5: Number of surveys per month

With our data, I achieved an average accuracy

of around 34 percent. I believe the low accuracy is due to the small sample size of only 140 documents, which makes it harder to accurately classify documents into taxonomy labels. Another challenge is that many documents have unique words in their titles and summaries, reducing the likelihood of word overlap between documents. Despite these limitations, I think the 34 percent accuracy reasonably reflects the taxonomy associations for each document.

2.4 Bonus

The bonus section extended the data evaluation phase by using two different approaches to analyze our data. The first approach was logistic regression, a machine learning algorithm suited for binary classification. It worked well with our normalized matrix and, due to its strength in predicting probabilities, achieved an average accuracy of around 41 percent. This was notably better than the RandomForestClassifier and highlighted the importance of trying different algorithms to understand why some yield higher accuracy, particularly as the size of the dataset (small or large) can influence performance.

LogisticRegression: 41.38%
Classification Report for LogisticRegression:

	precision	recall	f1-score	support
0	0.00	0.00	0.00	4
1	1.00	0.38	0.55	8
3	0.40	0.67	0.50	3
5	0.00	0.00	0.00	4
8	0.00	0.00	0.00	6
10	0.29	0.67	0.40	6
11	1.00	1.00	1.00	1
12	1.00	0.25	0.40	4
13	1.00	0.14	0.25	7
14	1.00	0.25	0.40	4
15	0.34	1.00	0.51	11
accuracy			0.41	58
macro avg	0.55	0.40	0.36	58
weighted avg	0.53	0.41	0.34	58

Figure 6: Number of surveys per month

The second approach involved using the RandomOverSampler technique, typically applied to imbalanced datasets. Although my dataset wasn't imbalanced, I decided to experiment with it to observe how it performs when its primary use case doesn't match the data. Surprisingly, it yielded a much higher accuracy of around 88 percent. This result was unexpected, but I believe part of the reason for the increased accuracy is that values previously set to 0 were adjusted closer to 1, helping balance the data. I reran logistic regression with the RandomOverSampler, and unlike in earlier evaluations, almost all documents had values associated with them.

RandomOverSampler Logistic Regression: 88.00%
Classification Report for LogisticRegression:

	precision	recall	f1-score	support
0	1.00	0.75	0.86	8
1	0.67	0.44	0.53	9
2	1.00	1.00	1.00	4
3	1.00	0.90	0.95	10
4	1.00	1.00	1.00	4
5	0.86	1.00	0.92	6
6	1.00	1.00	1.00	8
7	1.00	1.00	1.00	6
8	0.88	0.88	0.88	8
9	1.00	1.00	1.00	5
10	0.55	0.86	0.67	7
11	1.00	1.00	1.00	10
12	0.92	1.00	0.96	11
13	1.00	0.90	0.95	10
14	0.90	1.00	0.95	9
15	0.60	0.60	0.60	10
accuracy			0.88	125
macro avg	0.90	0.90	0.89	125
weighted avg	0.89	0.88	0.88	125

Figure 7: Number of surveys per month

3 Conclusion

In conclusion, there are various effective methods to analyze data and uncover trends, such as categorizing documents into their correct taxonomy labels. Techniques like logistic regression and RandomForestClassifier help assess accuracy, while simple Python libraries enable visualization through scatterplots and histograms. By manipulating data, we can track the number of surveys published per month and identify which categories have the highest publication volume

A APPENDIX

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*.