Exploring Large Language Model survey papers via Machine and Ensemble Learning

Mehenaz Afrin
Department of Computer Science
Boise State University
mehenazafrin@u.boisestate.edu

Abstract

Nowadays, there is an influx of researchers emphasizing Large Language Models (LLMs). While the field is broadening, it becomes difficult to keep up all the models, and techniques associated with the novel idea. To tackle this problem, a study has been conducted for assigning survey papers to taxonomy in an automated way. In this assignment, I am using their dataset for the task of exploration, manipulation, and evaluation. After finishing the instructed part, I did further exploration by using a cross tab between taxonomy and date, representing different visualizations for survey papers by taxonomy over time, and plotting the box of release day by taxonomy title. The experimental analysis indicates that Logistic Regression (LR) outperformed all the 8 Classifiers in terms of accuracy score, while GaussianNB (GNB) shows the most commendable precision score. For weighted recall and f1 score, LR shows the highest performance in text classification data.

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, in this project, I aim to explore, manipulate, and evaluate the existing dataset of Zhuang et al (Zhuang and Kennington, 2024)that consists of 144 survey papers for exploring LLMs survey paper with the help of Machine and ensemble learning techniques. I create a bar chart to visualize the survey paper per month in different taxonomies. I want to plot the release day by taxonomy title and figure out the relationship between date and taxonomy by using crosstab analysis. For feeding the dataset into models, I preprocess the dataset so that the machine can learn easily and combine all the features. 8 machine learning algorithms with ensemble methods have been implemented on the dataset for evaluation purposes. I did a detailed analysis by using three figures in the evaluation section. Overall, My contributions can be summarized as follows:

- Analyze the trends of survey papers over time, create a bar chart to see the number of published papers in different categories, and do a crosstab analysis to understand the relationship between taxonomy and date.
- Employ vectorizer, normalization techniques, and hot encoding technique to preprocess the data
- Several machine learning algorithms along with ensemble techniques have been applied for evaluation

2 Methodology

In this section, I will be doing data exploration, data manipulation, and data evaluation. The detailed analysis is discussed below.

2.1 Data Exploration

While the quantity of survey papers is massive enough, the instances are small compared to the survey papers. The dataset contains 16 categories of taxonomy along with their title, release date, summary, links, papered, author, and categories.

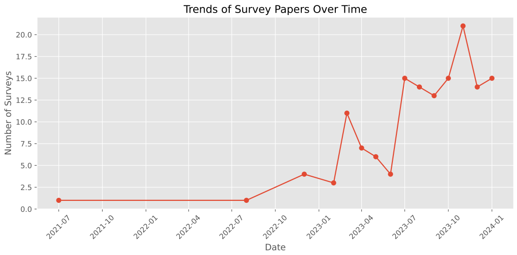

Figure 1: Trends of survey paper about large language models over time

After loading the dataset by using the pandas library, I try to analyze the trends of the survey paper over time. From Figure 1 we can see, that as time passes, the number of survey papers increases.

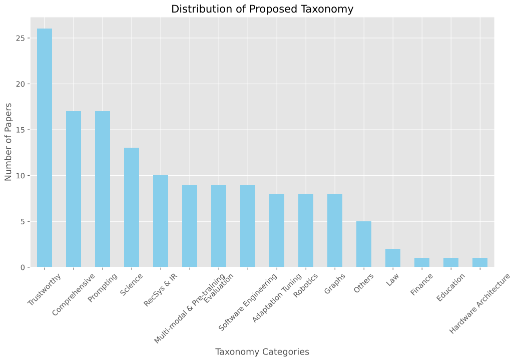

Figure 2: Distribution of survey papers across the taxonomy

Since the release date contains the date, month, and year, I group the number of survey papers by month and year then, count the survey papers by month and convert the data frame for facilitating the plotting of Figure 1. Here For styling the plots matplotlib applies ggplot which provides a professional and cleaner look. The mean value of the survey per month I get is 9.6 after calculating the mean value.

Initially, I tried to count the frequency of each category in the taxonomy for analyzing the distribution of the proposed taxonomy. I create a bar chart that shows the number of papers is assigned to each category in the taxonomy. The bar chart reveals that most survey papers are published in the trustworthy category, while fewer are found in finance and hardware architecture. The second most published paper was found in comprehensive and prompting. Software engineering, Multimodal and Pretraining, and Evaluation have a similar number of survey papers.26 survey papers are assigned to the trustworthy taxonomy.

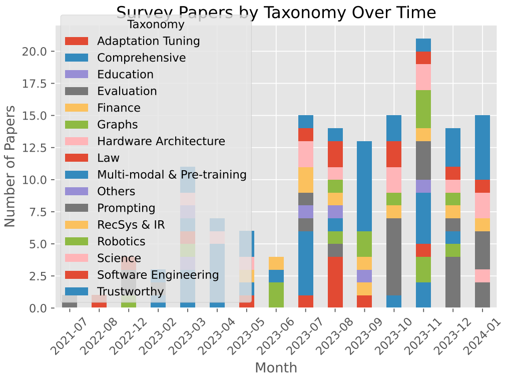

Figure 3: Survey paper by taxonomy over time

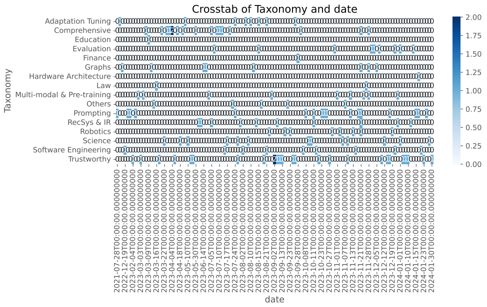

Figure 4: Correlation between taxonomy and date

Here Figure 4 analyzes the relationship between two categorical variables (taxonomy and release date) using crosstab. The boxplot displays a representation of the release day by taxonomy title.

2.2 Data Manipulation

In this section, I try to manipulate the dataset according to its nature. Since the dataset contains text values, I need to use a TF-IDF vectorizer for preprocessing. This vectorizer helps to retrieve the information precisely (Abubakar et al., 2022). I vectorize the title and summary column. Later I applied one hot encoding technique in the categories column. This method is useful for converting the categorical data into numerical values to feed the content to the machine since the machine only understands the numerical values (Seger, 2018). Later combining all of the features into a sparse matrix, thus building the feature matrix. Now when all the features are in numerical version, normalization technique has been applied to scale down features range between 0 to 1 and converge the features in a single unit form (Afrin et al., 2022). Identifying Taxonomy as the target column, I try to encode the label. Later, I allocate 40% dataset for testing and the rest 60% for training by setting the test size to 0.4.

2.3 Data Evaluation

I have employed 8 machine learning algorithms Support vector machine(Linear, Poly), GaussianNB (GNB), Decision tree (DT), k neighbors classifier (KNN), Logistic regression, XGB classifier, Random Forest (RF) classifier to evaluate the dataset in terms of accuracy, precision, recall, f score.

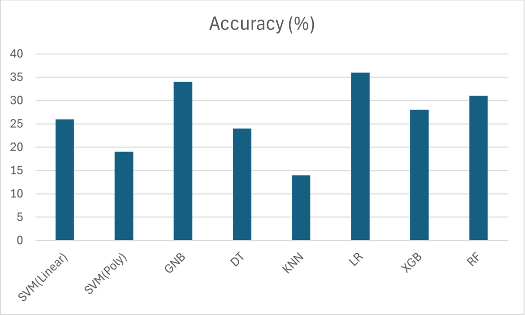

Figure 5: Performance level of all classifiers in terms of accuracy

Among all of the classifiers, Logistic regression outperforms in terms of accuracy and achieves 36% by dealing with the sparsity of the data. It works simply with interpretable coefficients indicating the intensity of each feature's influence on the outcome (Ifrim et al., 2008). Gaussian NB is well fit for TF-IDF text presentations by assuming Gaussian distribution for feature independence (Xu, 2018). While RF is powerful and robust, still GNB performs better than RF and shows 34% accuracy. GNB simplifies the calculation making it suitable for text data and leading to better performance. KNN performs worst among all of them by providing 14% accuracy.

Text data typically contains a large number of features (for example, words or tokens). KNN is based on distance calculations, which can be less useful in high-dimensional environments due to the "curse of dimensionality." As dimensions expand, data points get sparser, making it more difficult to discover meaningful neighbors (Moldagulova and Sulaiman, 2017). Its inherent limitations make it less fit for text classification compared to other methods. From Figure 5, we can see SVM linear gives more accuracy (26%) than SVM poly (19%) since text data contains linear separability characteristics. While polynomial kernels may represent noise, linear SVMs can efficiently capture these correlations without adding needless complexity. The decision boundaries of linear SVMs are simpler to understand compared to polynomial SVMs (Muflikhah and Haryanto, 2018). While DT provides 24% accuracy, XGBoost achieves 28% in terms of accuracy by emphasizing correcting errors going through sequentially tree construction. It integrates regularization and optimization for better performance than DT.

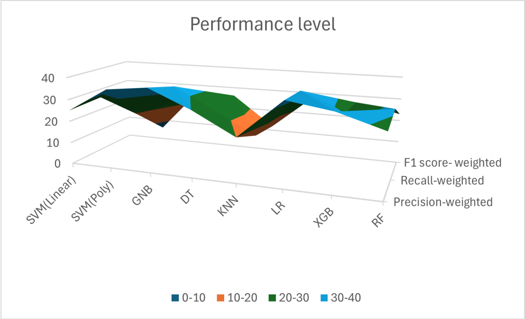

Figure 6: Performance level of all classifiers in terms of precision, recall, f score

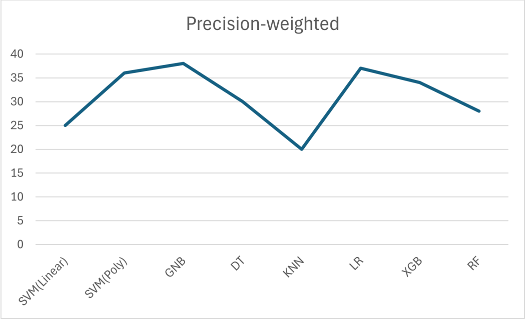

Figure 7: Performance level of all classifiers in terms of precision

The above figure demonstrates the weighted precision for each algorithm. While in terms of accuracy, LR performs best, GNB outperforms by achieving a 38% precision value. GNB discusses the strength of every independent feature through coefficients and provides impactful outcomes.LR shows 37% preciseness which is second best. XGB and DT achieve respectively 34% and 30% precision. Polynomial SVMs efficiently distinguish classes in datasets where nonlinearity exists among the features by generating complex decision boundaries. It enhances the precision by allowing the classifier to maintain against errors from non-support vectors. Therefore, polynomial SVM achieves a 36% precision score.

3 Conclusion

In this study, several machine learning algorithms along with ensemble learning have been applied to evaluate the performance level in terms of f score, precision, recall, and accuracy level. Data exploration has been done to visualize the fluctuation of survey papers over the period and to show the number of papers published in each 16 categories. Most of the papers are published in trustworthy taxonomy. Further, I analyze the correlation between date and taxonomy. The dataset has been processed by using normalization, TF-IDF vectorizer, and encoding techniques. In terms of precision GNB provides the best result while LR performed better in the remaining 3 evaluation metrices.

A APPENDIX

References

Haisal Dauda Abubakar, Mahmood Umar, and Muhammad Abdullahi Bakale. 2022. Sentiment classification: Review of text vectorization methods: Bag of words, tf-idf, word2vec and doc2vec. *SLU Journal of Science and Technology*, 4(1):27–33.

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*.

Mehenaz Afrin, Salma Akter Asma, Nazneen Akhter, Jaheed Hasan Ridoy, Sazida Sharmila Sauda, and Kazi Abu Taher. 2022. A hybrid approach to investigate anti-pattern from source code. In *2022 25th International Conference on Computer and Information Technology (ICCIT)*, pages 888–892. IEEE.

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.

Georgiana Ifrim, Gökhan Bakir, and Gerhard Weikum. 2008. Fast logistic regression for text categorization with variable-length n-grams. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 354–362.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Aiman Moldagulova and Rosnafisah Bte Sulaiman. 2017. Using knn algorithm for classification of textual documents. In *2017 8th international conference on information technology (ICIT)*, pages 665–671. IEEE.

Lailil Muflikhah and Dimas Joko Haryanto. 2018. High performance of polynomial kernel at svm algorithm for sentiment analysis. *Journal of Information Technology and Computer Science*, 3(2):194–201.

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.

Cedric Seger. 2018. An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing.

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.

Shuo Xu. 2018. Bayesian naïve bayes classifiers to text classification. *Journal of Information Science*, 44(1):48–59.

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*.