Taxonomy Classification of Large Language Model Survey Papers

Desmond Kofi Boateng
Department of Computer Science
Boise State University
desmondboateng@u.boisestate.edu

Abstract

This study develops a taxonomy classification model using traditional and deep learning methods. After data exploration, a Linear Support Vector Classifier (SVC) provides a baseline performance. However, a deep learning model with TF-IDF vectorization, dynamic learning rate scheduling, and early stopping proves more effective

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. New beginners commonly will read survey papers about LLMs to learn the recent progress in this field. Therefore, to facilitate their learning, numerous survey papers on LLMs have been published in the last two years. However, many 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 do an extensive exploration of the data to gain insights about the data and also work on manipulating the data into training and test data. We also aim to build a classifier model that is going to be trained on our data and will be used to classify the survey papers into taxonomies.

Overall, our contributions can be summarized as follows:

- Data exploration of the data
- Support Vector Machine Classifier
- Convolutional Neural Network Classifier

2 Related Work

Natural language processing (NLP) relies heavily on text classification, which finds wide-ranging applications in anything from news classification to screening scientific articles and classifying literature reviews. Text categorisation models are now much more efficient at managing massive amounts of data because to the development of machine learning and deep learning techniques. Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNNs), which are excellent at processing sequence data, have been made possible by advances in machine learning models like Naive Bayes and Support Vector Machines (SVM) as well as dictionary-based approaches. While LSTMs are well-suited for language problems because of its capacity to maintain consistency, CNNs, which were first created for image recognition, have been modified to capture spatial hierarchies in text.CNNs and LSTMs have shown effective in text categorisation, especially when handling subjective and objective datasets, as shown by Luan and Lin (Luan and Lin, 2019). Sunagar and colleagues (Sunagar et al., 2021) investigated machine learning techniques for classifying news topics through the use of algorithms such as Support Vector Machines (SVM), Naive Bayes, and K-Nearest Neighbours (KNN). The methods were assessed based on metrics including F1 score, accuracy, precision, and recall. Their research shown that machine learning approaches can produce remarkable outcomes in news categorisation tasks when combined with strong feature extraction and text pre-processing techniques. Text classification approaches have been effectively used for more specialised tasks including scientific article screening and literature survey analysis, in addition to news categorisation. SciBERT, a contextual language model pre-trained on scientific material, was used by Ambalavanan and Devarakonda (Ambalavanan and Devarakonda, 2020) to screen biomedical publications for systematic reviews. They contrasted ensemble designs, such as a cascade ensemble model, to a single integrated model known as the Individual Task Learner (ITL), framing the issue as a text classification task. Their results indicated that the cascade ensemble achieved improved precision, making it a better option for interactive search applications, although ITL was better suited for high-recall tasks. Scientific article screening saw a notable increase in classification accuracy when pre-trained language models like SciBERT were used. Additionally, McNabb and Laramee (McNabb and Laramee, 2017) conducted a "survey of survey papers" (SoS), grouping literature reviews according to themes, in order to address the difficulty of monitoring the expanding body of literature in information visualisation. Their analysis offered a well-organised framework to aid researchers in navigating the large number of survey papers, giving them insightful knowledge about both established and developing research fields. These developments in machine learning, especially the application of contextual language models and ensemble models, show how text classification is becoming more and more important in a variety of fields, including media, science, and literature management.

3 Methodology

3.1 Data Exploration

We start this by exploring the dataset given. We first perform an initial exploratory data analysis of the metadata of these papers. The dataset had some key attributes such as the title of the papers, the summary of the papers, the authors of the papers, the publication year and month, the arXiv links to the papers, and the arXiv IDs of the papers. These key attributes provide insights into the current area and trend of research in the area of LLMs.

We begin by exploring the distribution of the published papers over time. From 1, we observe that there has been a steady increase in the release of papers from July 2021 to January 2024. We also note that from July 2021 to December 2022, we see that the number of papers released during those months is quite low—about 4 papers released each month—but from the start of early 2023, we see a sharp rise in the number of publications. For example, in March 2023, there is a significant rise in the number of papers from about 3 papers in the previous month. The rise in the number of published papers continues till early 2024, when about 15 papers were released in January 2024. This trend shows a significant increase in LLM research, which is probably caused by the recent breakthroughs and success regarding generative pre-trained transformer (GPT) models.

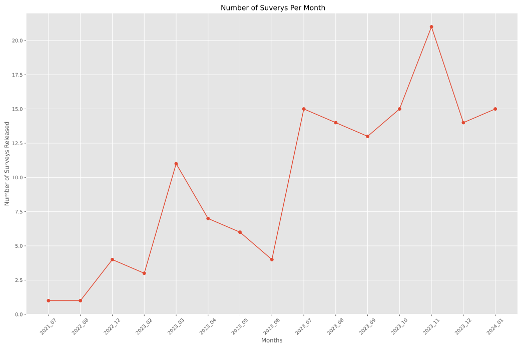

Figure 1: Number of Surveys Per Month

In Figure 2, we note the distribution of the papers across each taxonomy or topic. The most prominent taxonomy is "Trustworthy," which has about 26 papers released, followed by "Comprehensive" and "Prompting" with 17 papers each. We also note that taxonomy classes like "Law, Education, and Finance" are under-represented with about 3 papers each. This beckons that these are areas that have potential for further research and exploration.

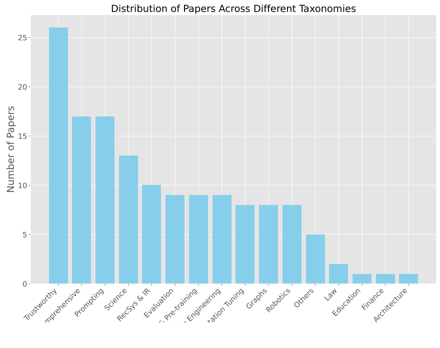

Figure 2: Distribution of Papers Across Different Taxonomies

We also observe the release patterns of papers across different taxonomies in Figure 3. We note that across most taxonomies there has been an upward trend in the release of papers from July 2021 to January 2024. A good example would be the "Comprehensive" taxonomy. We observe that in

2021 and 2022, there were no releases, but we note that there was a sharp increase in April and July 2023. We also note that from the "TrustWorthy" taxonomy there was a significant rise in the number of papers being released from March 2023 and peaked with 7 papers released in September 2023 and ending with 5 papers in January 2024. Similarly for "Prompting" and "MultiModal Pretraining", we see consistent releases starting in early 2023. This analysis shows that "Trustworthy" and "Comprehensive" have gained significant momentum in the last few months. This depicts that these areas of research have become increasingly important, and new research and development is being churned out on the regular.

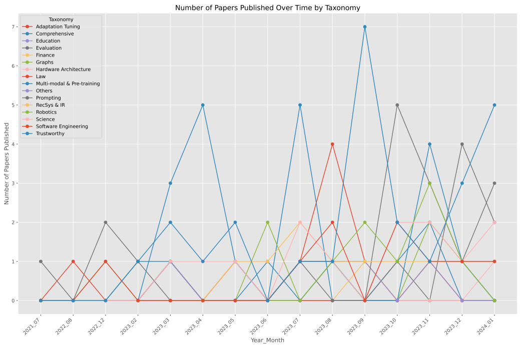

Figure 3: Distribution of Papers Across Different Taxonomies

3.2 Data Manipulation

In this section, we discuss the necessary steps taken to prepare the data to build a taxonomy classification model based on the dataset we have. The objective is to be able to convert the raw text and categories into a numerical format that's fit for most machine learning algorithms. This ensures that the machine learning model can learn effectively and make accurate predictions.
The first thing we do when we manipulate the data is to build a feature matrix, which serves as an input for the machine learning model to learn from. This is done by encapsulating the necessary features that we derived from the dataset. The first step in building a feature matrix is

1. **Vectorize text using TF-IDF.** The Term Frequency-Inverse Document Frequency(TF-IDF) is used to convert the text data from the title and summary columns in our dataset into a numerical data type. This captures the importance of words with respect to their frequency across the document. We start by creating two instances of the TF-IDF vectorizer to vectorize the 'Title" and 'Summary' columns. We set 'stop_words==english' to remove common English stop words from the analysis. We then use the fit_transform to learn the vocabulary of our dataset and transform it into TF-IDF matrices.

2. **One-hot Encoding of Categorical Variables** Next, we convert the 'Categories column into numerical format. We perform a one-hot encoding by splitting the categories column by the commas into multiple columns, and then we use pd.get_dummies to get binary columns for each category by representing whether they are present or absent in the categories for each paper.
   After this, we concatenate the TF-IDF matrices for the 'Title' column and "Summary' column with the one-hot encoding for the categories into one big feature matrix.

3. **Normalization** Next, we normalize the dataset to ensure that all variables contribute equally to the distance metric commonly used in machine learning algorithms. We use the Min-Max scaler from the sklearn.preprocessing module. We instantiated the Min-Max scaler, and then we transformed the feature matrix into a range of 0 to 1.

4. **Label Encoding** Next, we use label encoding to convert our target variable, which is the 'taxonomy' column, in a numerical format for our model. The label encoder converts each unique class in the taxonomy column into a numerical format that can be easily processed by our model.

5. **Train-Test Split** We used the train-test split to divide the normalized features into a training dataset and test dataset. The training dataset is to help our machine learning model learn from the data and then validate its performance on the test dataset. We set the TEST_RATIO=0.4. This means we keep 40% of our normalized features to test the accuracy of our model.

3.3 Data Evaluation

A Support Vector machine (SVM) is a supervised learning algorithm that is used mainly for classification and regression tasks. In the linear support vector machine, the goal is to find the best hyperplane that splits the data points into different categories with the maximum margin. In this case, our taxonomy categories are the classes that the hyperplane splits the data into the different classes. We trained the model on the training data, and we tested the accuracy of the model using the test data. We achieved an accuracy of 56.896%. We tried other models such as the random classifier, logistic regression, and naive bayes classifier, which performed better than the linear support vector machine.

3.4 Updating the Dataset

Next, we made a git pull of the GitHub link provided in Zhuang and Kennington (2024), and then we ran the scraping code to pull a lot more data on the survey papers. We scraped a total of 1924 papers, and after which we cleaned and dropped off all the papers that belonged to a different category other than the categories provided in the original dataset that was used to train the support vector machine.

After cleaning the data, we are left with a total number of 1050 rows. Once the data was cleaned, we built a neural network model to classify the papers in the new dataset into the 16b taxonomies. The model was built using a feed-forward convolutional neural network with TF-IDF for text vectorization and drop-out regularization to prevent overfitting. The target labels are derived from the taxonomy column of the dataset, and they are encoded using the LabelEncoder so that we can do a multi-class classification. We include both early-stopping and a variable learning rate scheduler. The early stopping criteria is to monitor the validation loss; if the validation loss does not change after 5 epochs, we stop the training. This is to prevent the model from overfitting. We also set a variable learning rate that reduces the learning rate when the validation loss plateaus. This allows the model to fine-tune its performance as the train continues. We evaluate the model on a validation dataset during training and also evaluate its performance on a new dataset, which is the test dataset. The model achieved a validation accuracy of 78.571% and testing accuracy of 73.33%.We run the model and predict the classes for all the rows of the scraped data. In order to ensure that the updated dataset had accurate classification, we dropped all classifications whose probabilities were less than 0.9. In the end, we have a total dataset of about 250 rows of data with papers that have been classified.

4 Conclusion

The models developed in this paper after deep exploration into the data and an attempt to understand the data and the data types. This helped to explore and manipulate the data, preparing it for the model-building process. We initially experimented with a linear support vector machine (SVM). The support vector machine gave a reasonable accuracy of 56.896%. While the accuracy was okay, the SVM was limited; hence, we had to build a better model. Next we built a convolutional neural network (CNN), which improved the testing accuracy to 73.33%. This was better at classifying the survey papers than the SVM did. We also ran the scraping scripts to update the datasets, and we ran the random classifier on it to update the dataset and add new data to it.

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*.

Ashwin Karthik Ambalavanan and Murthy V Devarakonda. 2020. Using the contextual language model bert for multi-criteria classification of scientific articles. *Journal of biomedical informatics*, 112:103578.

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*.

Yuandong Luan and Shaofu Lin. 2019. Research on text classification based on cnn and lstm. In *2019 IEEE international conference on artificial intelligence and computer applications (ICAICA)*, pages 352–355. IEEE.

Liam McNabb and Robert S Laramee. 2017. Survey of surveys (sos)-mapping the landscape of survey papers in information visualization. In *computer graphics forum*, volume 36, pages 589–617. Wiley Online Library.

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.

Pramod Sunagar, Anita Kanavalli, Sushmitha S Nayak, Shriya Raj Mahan, Saurabh Prasad, and Shiv Prasad. 2021. News topic classification using machine learning techniques. In *International Conference on Communication, Computing and Electronics Systems: Proceedings of ICCCES 2020*, pages 461–474. Springer.

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*.