Classifying of research papers based on their research areas or topics using machine learning

NAMYALO AGNES
Department of Computer Science
Boise State University
agnesnamyalo@u.boisestate.edu

Abstract

With the rapid advancement of computer and information technologies, a vast number of research papers are now available both online and offline. As new research fields continue to emerge, users face significant challenges in finding and categorizing papers of interest. To address these issues, this paper presents a research paper classification system that groups papers into meaningful categories based on shared topics. The system identifies representative keywords from each paper and uses the Term Frequency-Inverse Document Frequency (TF-IDF) method to measure the importance of words and measure of how many times a word appears in a document. The logistic regression algorithm is then employed to classify papers with similar topics.

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, we aim to classify the survey papers into different categories (topics) using logistic regression.

Overall, our contributions can be summarized as follows:

- Importing data
- Data exploration
- Preprocessing of data
- Machine learning

2 Related Work

A number of classifcation techniques have been used for document classification. (Barigou, 2018)divides Automatic document classifcation into two methods: supervised and unsupervised. (Pradeepa et al., 2024) proposes logistic regression for text classification and identifies two main advantages of logistic regression where it can naturally provide probabilities and extend to multi-class classifcation problems Another advantage is that most of the methods used in logistic regression model analysis follow the same principles used in linear regression. (Shah et al., 2020) have compared logistic regression, random forest and K-nearest neighbour as classification algorithms for BBC news text classification and the authors decided to show the comparison based on five parameters namely precision, accuracy, F1-score, support and confusion matrix.However, (Indra et al., 2016) specifically considers logistic regression to classify tweets into the selected topics.

3 Methodology

3.1 Data Exploration

The research was conducted using a dataset of 144 different survey papers with 8 attributes as shown in Table 1. For these survey papers, a taxonomy was designed where each paper was assigned to a corresponding category within the taxonomy. To better understand the data, a line graph as illustrated in Figure 1 which shows that the survey papers in the

dataset range from July 2021 to January 2024 and the number of survey papers has been increasing significantly.

Furthermore, a bar graph for the taxonomy attribute was used to better understand the distribution of the classes as shown in Figure 2.With the data being rightly skewed and having more of the taxonomy of papers being Trustworthy(with 26 papers) The distribution indicates that the class is extremely imbalanced.Further more data distribution was analysed using a bar graph in Figure 3 which shows the distribution of survey papers across different categories with cs.CL being the most frequent category authors choose for their works.

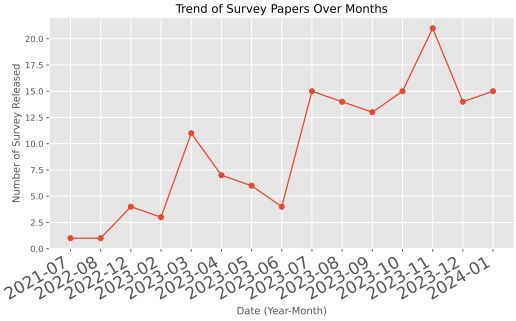

Figure 1: Line graph to showing trends of survey papers over months

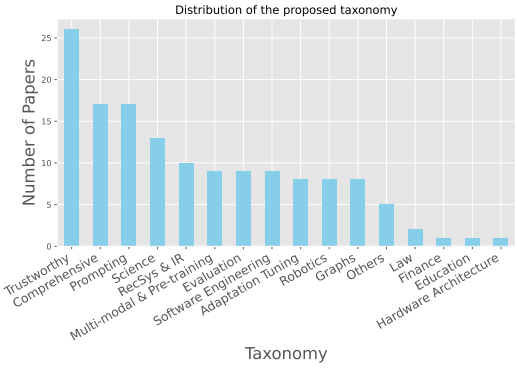

Figure 2: Distribution of the proposed taxonomy

3.2 Data Manipulation

The data was manipulated to represent text data as a numerical matrix where feature matrices for both title and summary were created using the term frequency-inverse document frequency (TF-IDF). According to (Kim and Gil, 2019) describes TF-IDF as a technique used extensively to retrieve information and mine text to assess the significance of each word within a collection of documents.

After establishing a feature matrix, we went

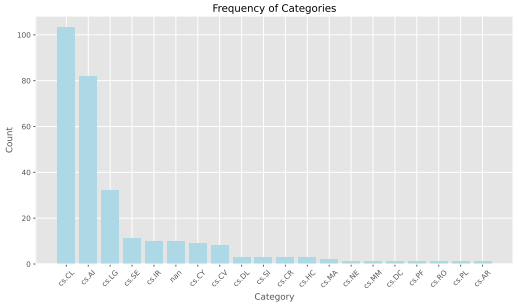

Figure 3: Frequency of Categories

further to preprocess the feature matrix which involved preparing the data to be in the right format for the machine learning models, the following activities were applied in the preprocessing of the feature matrix; normalizing the data, one hot encoding, encoding. Further more, the dataset to be used for evaluation was split into two parts: training and testing sets using a test size of 0.4.The splitting of data enabled the model to be trained on one part of the data (the training set) and evaluate its performance on unseen data (the testing set).

3.3 Data Evaluation

Attribute	Description
Taxonomy	Proposed taxonomy
Title	Paper title
Summary	Abstract of papers
Authors	Lists of author's name
Links	Links of papers
Paper ID	Paper ID
Categories	Category
Release Date	First released date

Table 1: Data attributes and description.

A logistic regression classifier was employed and trained on the model and predicted on the test set and parameters namely accuracy, precision, F1-score was given with an accuracy of 0.22. Having noticed that the accuracy was very low, we went further to scale the data and handle the imbalanced data since one class appeared to have more instances than another significantly. The technique of assigning class-weight to balanced was used and helped the model to pay more attention to the minority class.The accuracy was increased to 0.47. Furthermore, we went further to apply hyperparameter tuning to optimise the parameters of the model and this led to an improved accuracy of 0.52.

The results for the logistic regression with the

hyper-tuned parameters are shown in Figure 4 with the performance metrics where accuracy of 0.52 is obtained. This figure shows that all the parameters for each of the individual classes are calculated.

Other machine learning algorithms like SVM and Random Forest for comparison were also run but still gave lower accuracies and not considered.

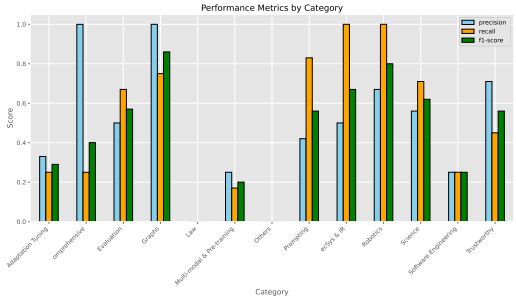

Figure 4: Graph showing the metrics of Logistic regression

3.4 Conclusion, challenges and future work

According to the logistic regression algorithm, most research papers appear to be from graphs research topics. However, a challenge of low accuracy was encountered even after handling class imbalance. For future works, this can be handled by employing more machine learning algorithms for classification for comparison with the logistic regression in order to have improved evaluation metrics.

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*.

Fatiha Barigou. 2018. Impact of instance selection on knn-based text categorization. *Journal of Information Processing Systems*, 14(2):418–434.

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.

ST Indra, Liza Wikarsa, and Rinaldo Turang. 2016. Using logistic regression method to classify tweets into the selected topics. In *2016 international conference on advanced computer science and information systems (icacsis)*, pages 385–390. IEEE.

SW Kim and JM Gil. 2019. Research paper classification systems based on tf-idf and lda schemes. human-centric computing and information sciences.

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

S Pradeepa, Elizabeth Jomy, S Vimal, Md Mehedi Hassan, Gaurav Dhiman, Asif Karim, and Dongwann Kang. 2024. Hgatt_lr: transforming review text classification with hypergraphs attention layer and logistic regression. *Scientific Reports*, 14(1):19614.

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.

Kanish Shah, Henil Patel, Devanshi Sanghvi, and Manan Shah. 2020. A comparative analysis of logistic regression, random forest and knn models for the text classification. *Augmented Human Research*, 5(1):12.

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*.