A Supervised Approach to Zero-Shot Learning for Field Classification of Texts: Leveraging File Data for Improved Text Categorization

Krishna Advaith Siddhartha Rangavajjula
Project Intern at CSIR – IICT, Student at MVSR Engineering College, Hyderabad, Telangana, India

Anil Kumar Pulipaka
Principal Scientist, Business Development & Research Management (BDRM), CSIR-Indian Institute of Chemical Technology, Hyderabad, Telangana, India

ABSTRACT

Assessing work from various fields is necessary to analyze & survey an institution's performance over a certain period of time having progression in multiple divisions. Many necessary characteristics such as Impact Factor (IF) of acclaimed works are influenced by non-uniform distribution of publications in different sections and renowned journals. Classifying file elements with NLP based on the publication titles would be supportive and intuitive. Text Analysis and Field Classification requires a large amount of data for a model to be trained and efficient. So, a Zero shot learning approach is opted to distinguish various publications into their respective sectors. Unlike other models, this model is enhanced to leverage CSV format files for both input and output. Different Pre-Trained Language models have been used and their performances are recorded. The advantage of zero shot learning over regular methods is discussed.

General Terms

Natural Language Processing, Pre-Trained Language Model, Text Classification, Transformers, Zero Shot Learning.

Keywords

Natural Language Processing, Pre-Trained Language Model, Text Classification, Transformers, Zero Shot Learning.

1. INTRODUCTION

In this growing world, having a clear concise picture on developments would help any institution to analyze & understand its contributions to several distinct categories. This would support their thought of reacting wisely over the next period based on the previous records and data.

To classify each section of work done, lots of human power is consumed. The use of machine learning could help the required work to be done faster and simpler, by building an automated classifier. Classification tasks require various prerequisites such as heavy amounts of training and testing data, a standardized model that executes with satisfactory accuracy scores. This would require input data to be stored in a systematic ordered structure. Whereas accessing children's data directly from files would help users i.e., the members of the institution.

NLP has become a crucial technology in today's world, transforming the way people communicate with data and technology. Sophisticated NLP applications are needed in various fields like customer service, healthcare, finance, and other sectors. Text categorization, a crucial aspect of NLP, involves the classification of text into predetermined categories. This task is crucial for purposes such as sentiment analysis, spam detection, and topic categorization. With the increase in size and intricacy of text data, conventional text classification techniques frequently face difficulty in keeping up. This has resulted in the development of zero-shot learning (ZSL), a concept that allows models to categorize data into classes that were not encountered during training. ZSL utilizes pre-trained language models (LMs) that have been trained on extensive data and show a deep comprehension of language. Pre-trained language models like BERT, GPT-3, and RoBERTa have demonstrated exceptional results in a range of NLP assignments. They offer a strong base for incorporating ZSL, as they have the ability to effectively adapt to unfamiliar categories due to their thorough pre-training. This feature is especially useful in ever-changing settings where new categories often appear, and labeled data is not accessible for all potential classes. Assessing various ZSL models entails examining how well they perform on different factors like accuracy, generalization capability, and computational efficiency.

Having a grasp of these distinctions is essential when choosing the best model for a particular use case. In this opening, the importance of NLP and text classification, examine the idea and advantages of zero-shot learning, talk about the significance of pre-trained language models, and offer a comparison of different ZSL models are discussed. This investigation will showcase the progress in the industry and lead the way.

2. WHY IS ZERO SHOT LEARNING PREFERRED OVER REGULAR MACHINE LEARNING ?

Zero-shot learning (ZSL) is a cutting-edge paradigm in machine learning and natural language processing (NLP) that enables models to perform tasks without having been explicitly trained on specific examples for those tasks. Unlike traditional supervised learning, which relies on large amounts of labeled data, zero-shot learning leverages the ability of models to generalize from a wide array of data and tasks they have seen during their training phase.

At its core, zero-shot learning hinges on the use of pre-trained language models that have been exposed to vast amounts of text data from diverse sources. These models, such as GPT-3, T5, RoBERTa, and others, are trained on massive datasets encompassing a wide variety of topics and contexts [1][2][3]. This extensive pre-training enables them to develop a deep understanding of language and its nuances. Zero-shot learning typically involves three main steps: pre-training, task definition, and inference. The model is trained on a large corpus of text data, learning to understand and generate human language. When a specific task is presented, such as text classification, sentiment analysis, or question answering, the task is defined in terms of natural language prompts or descriptions. The model then uses its pre-trained knowledge to infer the correct label or perform the task based on the given prompts, without having seen any task-specific labeled data during training.

Zero-shot learning offers several compelling advantages that make it a preferred choice for many applications. Traditional machine learning models require large labeled datasets, which are expensive and time-consuming to create. Zero-shot learning bypasses this need by leveraging pre-existing knowledge embedded in the model. This is particularly useful in domains where labeled data is scarce or difficult to obtain [3][4]. Zero-shot models are highly versatile and can be quickly adapted to new tasks or domains without the need for retraining. This flexibility is invaluable in dynamic environments where new tasks emerge frequently, and the ability to respond quickly is crucial [1][5].This makes it accessible for organizations with limited resources [6]. Zero-shot learning models can handle a wide range of tasks simultaneously. For instance, a single model can perform text classification, translation, summarization, and more, simply by providing appropriate prompts. This scalability is a significant advantage for applications that require multitasking capabilities [3]. The ability to perform tasks out-of-the-box enables rapid prototyping and deployment of AI solutions. Developers can test ideas and iterate quickly without the lengthy process of collecting and labeling data [5][6]. Pre-trained models used in zero-shot learning have typically seen a vast and diverse range of data. This broad exposure allows them to generalize better across different tasks and domains compared to models trained on limited datasets [1][2].

Zero-shot learning is being increasingly adopted across various industries and applications. It is used for text classification, categorizing documents, emails, or any text data into predefined categories without needing labeled examples for each category [3][4]. For sentiment analysis, it determines the sentiment of text (positive, negative, neutral) even when specific examples of sentiment-labeled data are not provided [1]. It enhances search engines by allowing them to understand and retrieve relevant information based on context and query understanding without needing specific training for each type of query [4]. Zero-shot learning is also employed in named entity recognition (NER), identifying and classifying entities (like names, dates, places) in text without being explicitly trained on annotated NER datasets [6]. Furthermore, it is utilized in machine translation and language understanding, translating text between languages and understanding multilingual content, leveraging the pre-trained multilingual capabilities of models like mBERT [2].

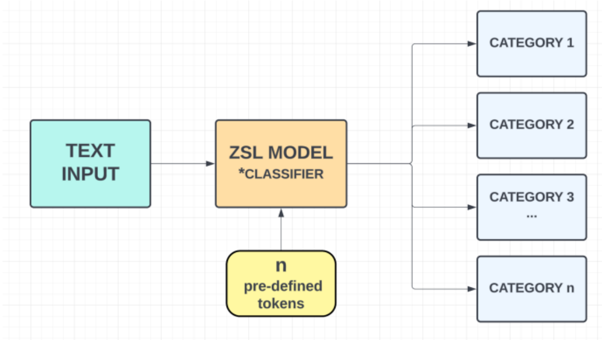


Fig 1: Zero shot text classification demonstration

The figure above is a simple demonstration of how zeroshot text classification works.

Here, a list of labels/categories are provided in which the user inputs are expected to fall under, and respective scores for each category is expected in return. Category with the highest of those scores is considered to be the final result. For example, given three statements are example inputs and had set our labels as “Interrogative”, “Declarative” and “Exclamatory”. And here are our scores for those three samples:-

- **“Hari is writing a book”:** This is selected to be a Declaratory sentence with 52.3% confidence.
- **“Visali asked a question”:** Even though it has the word ‘question’, this sample is selected to be a “Declaratory” sentence with 47.9% confidence.
- **“Apoorva said Great!”:** This sentence is selected to be under “Exclamatory ” with a confidence of 77.8% .

Zero-shot learning represents a significant advancement in the field of artificial intelligence, providing a robust and efficient way to tackle a myriad of tasks without the traditional reliance on extensive labeled datasets. Its ability to generalize from broad pre-training, combined with its flexibility and cost-effectiveness, makes it an attractive option for researchers and practitioners alike. As AI continues to evolve, zero-shot learning is poised to play a pivotal role in driving innovation and expanding the horizons of what is possible with machine learning.

3. ANALYSIS AND SELECTION OF VARIOUS PRE-TRAINED LANGUAGE MODELS

Pre-trained Large Language Models (LLMs) are a major breakthrough in artificial intelligence, utilizing extensive data and complex structures to comprehend and produce text similar to human language.. The initial training stage includes handling extensive data to grasp grammar, information, and context, which can later be adjusted for particular tasks such as translation, summarization, or answering questions. The flexibility and expandability of pre-trained LLMs have transformed natural language processing, becoming highly valuable in a wide range of uses from customer service chatbots to cutting-edge research tools, thus expanding the possibilities of AI in comprehending and producing human language. Here are some of the famous pre trained language models that are capable of text classification with high accuracy scores :

3.1 BERT

BERT (Bidirectional Encoder Representations from Transformers) is a cutting-edge NLP model created by Google AI in 2018. In contrast to conventional models, BERT trains bidirectionally by analyzing text in both directions at the same time, improving contextual comprehension. The Transformer architecture relies on its encoder for its operation.

3.2 GPT-3

GPT-3, which stands for Generative Pre-trained Transformer 3, is a sophisticated language model created by OpenAI. Utilizing 175 billion parameters, it creates text similar to that of humans by utilizing input prompts. GPT-3 performs well in many activities, such as translation, summarization, and question answering, thanks to its extensive pre-training on a wide range of internet text.

3.3RoBERTa

RoBERTa, short for Robustly optimized BERT Approach, is a natural language processing model created by Facebook AI. It enhances BERT through training on increased data and using bigger mini-batches. RoBERTa eliminates the Next Sentence Prediction target found in BERT and places emphasis on dynamic masking and extended training

3.4 ALBERT

ALBERT (A Lite BERT) is a BERT model variation created for improved efficiency and scalability. Created by Google

Research, ALBERT decreases model size by using factorized embedding parameterization and sharing parameters across layers.
In addition to these, XLNet (Generalized Autoregressive Pre Training for Language Understanding) and CLIP (Contrastive Language-Image Pre-training) are tested under 3 classification problems including our tokenized field classification. The classification problems and their recorded results are :

In Text Categorization, models are required to classify news article titles into respective sections of news they belong to with their respective scores. AG News Classification Dataset is used for this process and it is observed that GPT-3 has shown its superiority over the other models when accuracy is considered with 85 - 90%. Whereas, RoBERTa stands second in performance by 80 - 90% scores followed by XlNet , BERT and ALBERT. But, when time taken by each of the models for completion is considered, ALBERT and BERT are proven to be much faster, giving the next positions to RoBERTa and XLNet. GPT-3 was surprisingly time-taking which took twice the amount of time in comparison with its successive model. Whereas in Spam Detection, Enron-Spam dataset is used and the results were following the pattern from previous record, with GPT-3 being the most accurate and time taking, and RoBERTa being a better choice satisfying both speed and performance.

In Field Classification, text samples from publication records are provided as test input, with various fields of science as the labels. Consolidating previous result analysis, the test models have performed in the exact same behavior as regular text categorization. Therefore, GPT-3, RoBERTa and BERT models are taken under consideration for further proceedings. And after observation, RoBERTa is selected to be the appropriate model for our application.

Table – 1 : Comparison of Zero Shot Models over Several Classification Problems

ORDER OF PREFERENCE	MODEL	TEXT CATEGORIZATION	AVG TIME	SPAM DETECTION	AVG TIME	FIELD CLASSIFICATION	AVG TIME
4	BERT	75-85%	60-200 ms	88-92%	50-150 ms	75-85%	60-200 ms
5	XLNet	80-88%	90-260 ms	89-93%	80-220 ms	80-88%	90-260 ms
1	**RoBERTa**	**80-90%**	**100-250 ms**	**90-94%**	**80-200 ms**	**80-90%**	**100-250 ms**
2	GPT-3	85-90%	250-500 ms	94-96%	200-400 ms	85-90%	250-500 ms
3	ALBERT	75-85%	50-160 ms	87-91%	40-120 ms	75-85%	50-160 ms
6	CLIP	70-80%	350-700 ms	85-89%	300-600 ms	70-80%	350-700 ms

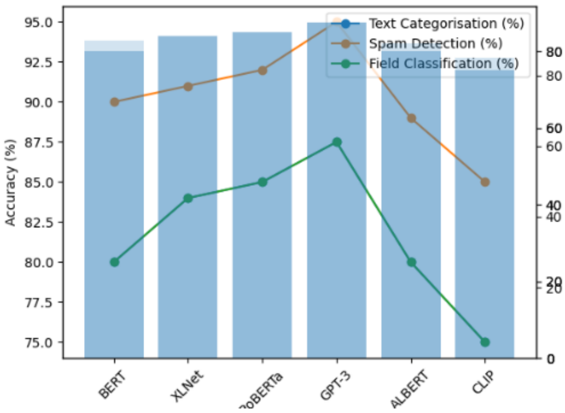


Fig 2: Accuracy comparison of models

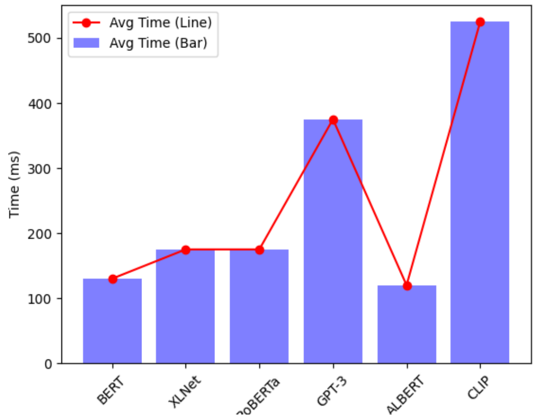


Fig 3: Speed Comparison of models : average time taken by each of the Models

3.5 WHY IS ROBERTA PREFERRED OVER OTHER ZERO-SHOT MODELS?

- **Efficacy and Resource Requirements:** RoBERTa promises better understanding, reactiveness and accurate language processing of input data.
- **Data for training and domain adaptability:** RoBERTa is trained on a much larger database than the rest of other models.
- **Cost and Accessibility:** Due to its smaller size and lower computational requirements, RoBERTa is more cost-effective to use in production environments.
- **Balance in Performance and Speed:** RoBERTa promised a better balanced result in both Performance and Speed , as shown in the above table.
- **Can be Fine-Tuned for Specific Tasks:** RoBERTa excels in tasks that require deep understanding and nuanced handling of text due to its pre-training on a large, diverse corpus.

4. RELATED WORK

CSIR-Indian Institute of Chemical Technology(CSIR-IICT) - a constituent laboratory under Council of Scientific & Industrial Research (CSIR) annually publishes reports showcasing its research & development achievements and publications. These reports encompass a wide range of topics across varied scientific fields. However, they are not classified under selected categories such as Chemical, Healthcare, Energy and Environment..etc.,. The lack of categorization makes it hard for the Research & Development Department to assess the institution's influence in specific research areas of interest.

Traditional methods for text categorization require extensive labeled data for training which can be expensive and time-consuming to generate. However, zero-shot text classification offers a promising alternative by utilizing pre-trained language models to classify text without requiring labeled examples for each specific class. This approach is particularly advantageous for R & D institutions like CSIR-IICT where it is tedious to manually tag numerous documents.

This study uses zero-shot text classification models to automatically categorize annual reports from institutions. Our objective is to provide a brief and structured overview of institution's influence in selected research areas. The process of categorization involves using advanced models (such as RoBERTa) which have shown to be useful in various natural language processing tasks. These models are subjected to fine-tuning and evaluation to ensure high accuracy in categorizing documents.

Utilizing zero-shot text classification enhances the effectiveness of organizing and analyzing research results as well as empowers users to make informed decisions with categorized data. This research highlights how zero-shot learning models are effective in handling the classification of complex and diverse research papers which enhances data analysis and reporting in research and development institutions. The results also show the actual benefit of zero shot learning methods compared to traditional approaches that necessitate numerous rounds.

Input File Structure:

1. **S.No**
2. **Author of the Publication**
3. **Title of Publication**
4. **Title of the Journal**
5. **Volume**
6. **Issue**
7. **Month of Publishing**
8. **Page . no**
9. **Impact Factor**
10. **Division**

Currently, The "Title of Publication" is used as classifier input and key of the model.

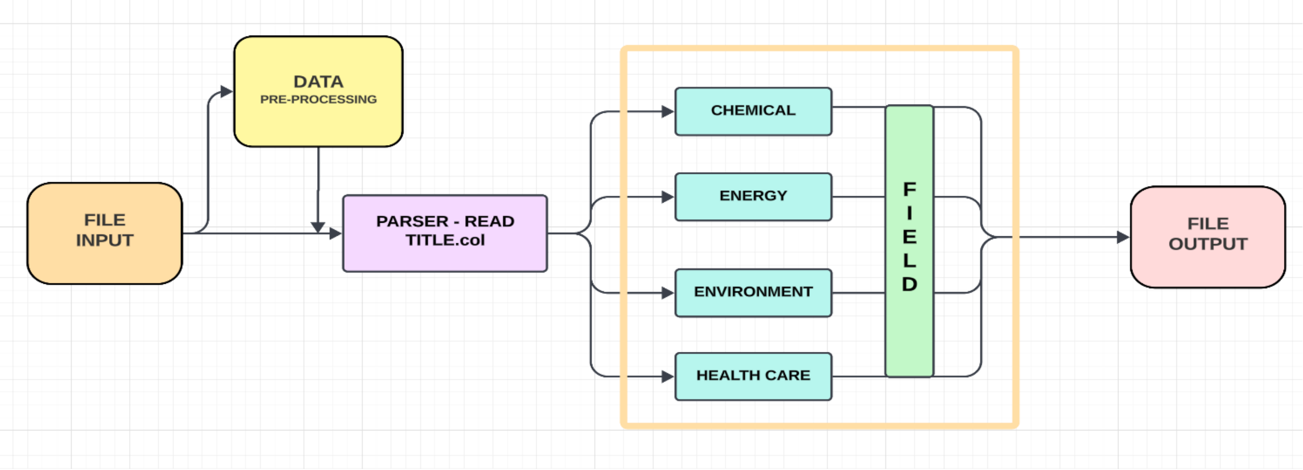


Fig 4: Work Flow Diagram

5. STAGES OF WORKING AND EVALUATION

1) **File input:** Instead of ordered text samples, a '**.csv**' file having details about publications by the Research & Development Department. It Consists of Title, Impact Factor, Author details and Journal it was Published in, Volume etc.,. This data can be Previewed, sorted and Downloaded with the help of streamlit's display option.

2) **Data Pre-Processing:** Pre-processing is a crucial step in preparing the raw data for further analysis. This involves cleaning and structuring the data to make it suitable for classification.

3) **File read:** The parser specifically reads the titles of the publications which are assumed to be in a column labeled as "TITLE" and Title of each publication is Extracted.

4) **Classifier:** This step involves classifying the titles into various predefined categories. Multi-potential categories are considered for each title. The pre-processed title is passed through a zero-shot text classification model. The model evaluates likelihood of the title belonging to several predefined categories.

5) **Selection:** Out of all possible categories suggested by the model, one with highest confidence score is selected.

6) **Assigning Fields:** Once, the most appropriate category is selected, it is assigned to publication. This step finalizes the classification process.

7) **File Output:** Once, one assigns the Fields into file, exact file is returned as output for further analysis, review and observations.

The Output file can be further reviewed, Sorted, Classified and Downloaded through Streamlit.

5.1 Evaluating The Performance of a Zero Shot Learning Model :

To ensure that the results would be same throughout the process, Program is executed a couple of times with a similar data frame of 50 &100 randomly chosen indices and analyzed the scores and final result. All of our reruns have produced the exact result without any kind of loss from the original data that has been given.

Considering a real-time dataset of annual publications From CSIR – IICT as our sample source. After multiple reruns with data arrangement permutations, results are observed to be consistent.

Table-2 : Results of 50-Dataset

RERUN	DATA SET	TIME (in seconds)	ACCURACY	ERROR (%)	DELTA ERROR	RESULT	RESULT ANALYSIS
0	50- KNOWN VALUES SET - 1	56.95	HUMANE	0%		Chemical: 43 Environment: 5 Energy: 2	SET – 1 RECORDED
1	50- KNOWN VALUES SET - 2	54.51	100%	0%	0	Chemical: 43 Environment: 5 Energy: 2	ZERO CHANGE (SET 1, SET 2)
2	50- KNOWN VALUES SET - 3	49.81	100%	0%	0	Chemical: 43 Environment: 5 Energy: 2	ZERO CHANGE (SET 1, SET 2, SET 3)
3	50- KNOWN VALUES SET - 4	55.13	100%	0%	0	Chemical: 43 Environment: 5 Energy: 2	ZERO CHANGE (SET 1, SET 2, SET 3, SET 4)
4	50- KNOWN VALUES SET – 5	50.26	100%	0%	0	Chemical: 43 Environment: 5 Energy: 2	ZERO CHANGE (SET 1, SET 2, SET 3, SET 4, SET 5)
5	50- KNOWN VALUES SET - 6	48.76	100%	0%	0	Chemical: 43 Environment: 5 Energy: 2	ZERO CHANGE (SET 1, SET 2, SET 3, SET 4, SET5, SET 6)

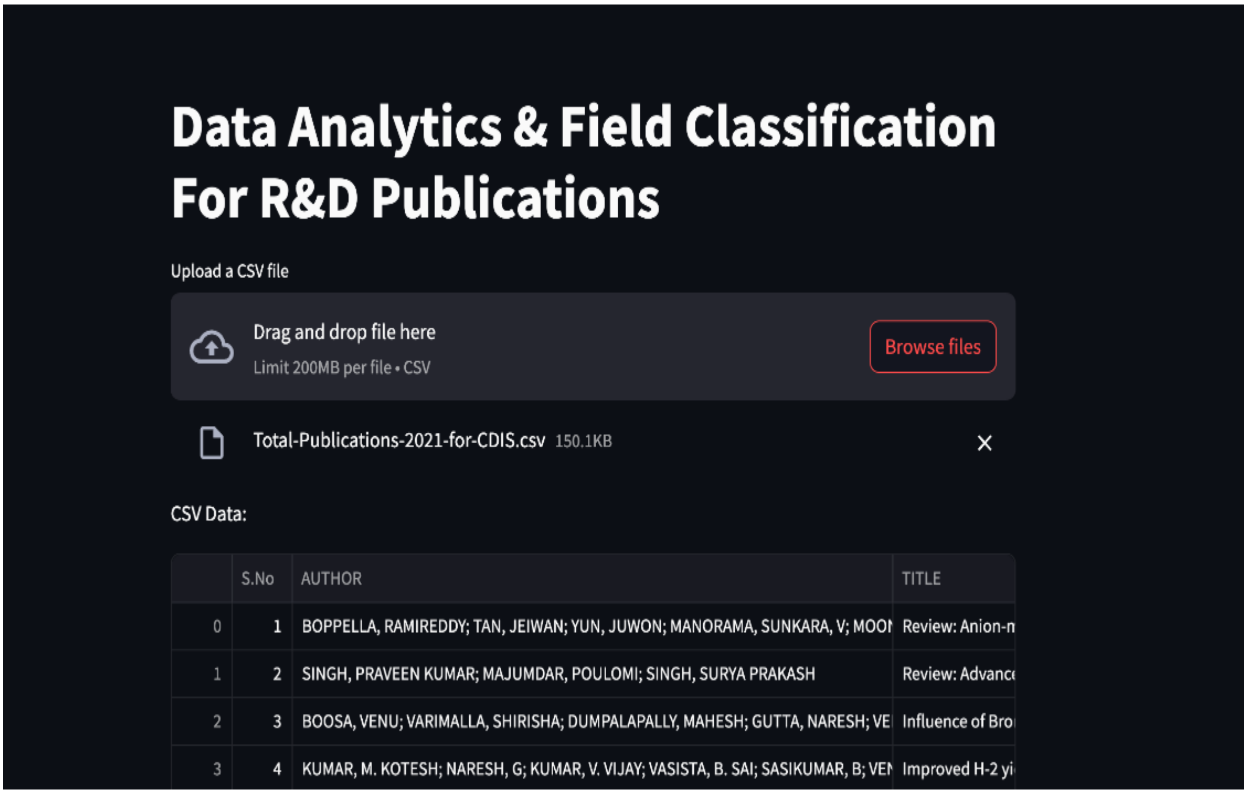
Table-3 : Results of 100-Dataset:

RERUN	DATA SET	TIME (in seconds)	ACCURACY	ERROR (%)	DELTA ERROR	RESULT	RESULT ANALYSIS
0	100- KNOWN VALUES SET - 1	107	HUMANE	0%		Chemical: 75 Health care: 5 Environment: 16 Energy: 4	SET – 1 RECORDED
1	100- KNOWN VALUES SET - 2	104	100%	0%	0	Chemical: 75 Health care: 5 Environment: 16 Energy: 4	ZERO CHANGE (SET 1, SET 2)
2	100- KNOWN VALUES SET - 3	109.81	100%	0%	0	Chemical: 75 Health care: 5 Environment: 16 Energy: 4	ZERO CHANGE (SET 1, SET 2, SET 3)
3	100- KNOWN	107.81	100%	0%	0	Chemical: 75 Health care: 5 Environment: 16	ZERO CHANGE (SET

	VALUES SET - 4					Energy: 4	1, SET 2, SET 3, SET 4)
4	100-KNOWN VALUES SET – 5	105	100%	0%	0	Chemical: 75 Health care: 5 Environment: 16 Energy: 4	ZERO CHANGE (SET 1, SET 2, SET 3, SET 4, SET 5)
5	100-KNOWN VALUES SET - 6	106.98	100%	0%	0	Chemical: 75 Health care: 5 Environment: 16 Energy: 4	ZERO CHANGE (SET 1, SET 2, SET 3, SET 4, SET5, SET 6)

We can conclude that the results are consistent throughout the process.

6. RESULTS



Data Analytics & Field Classification For R&D Publications

Upload a CSV file

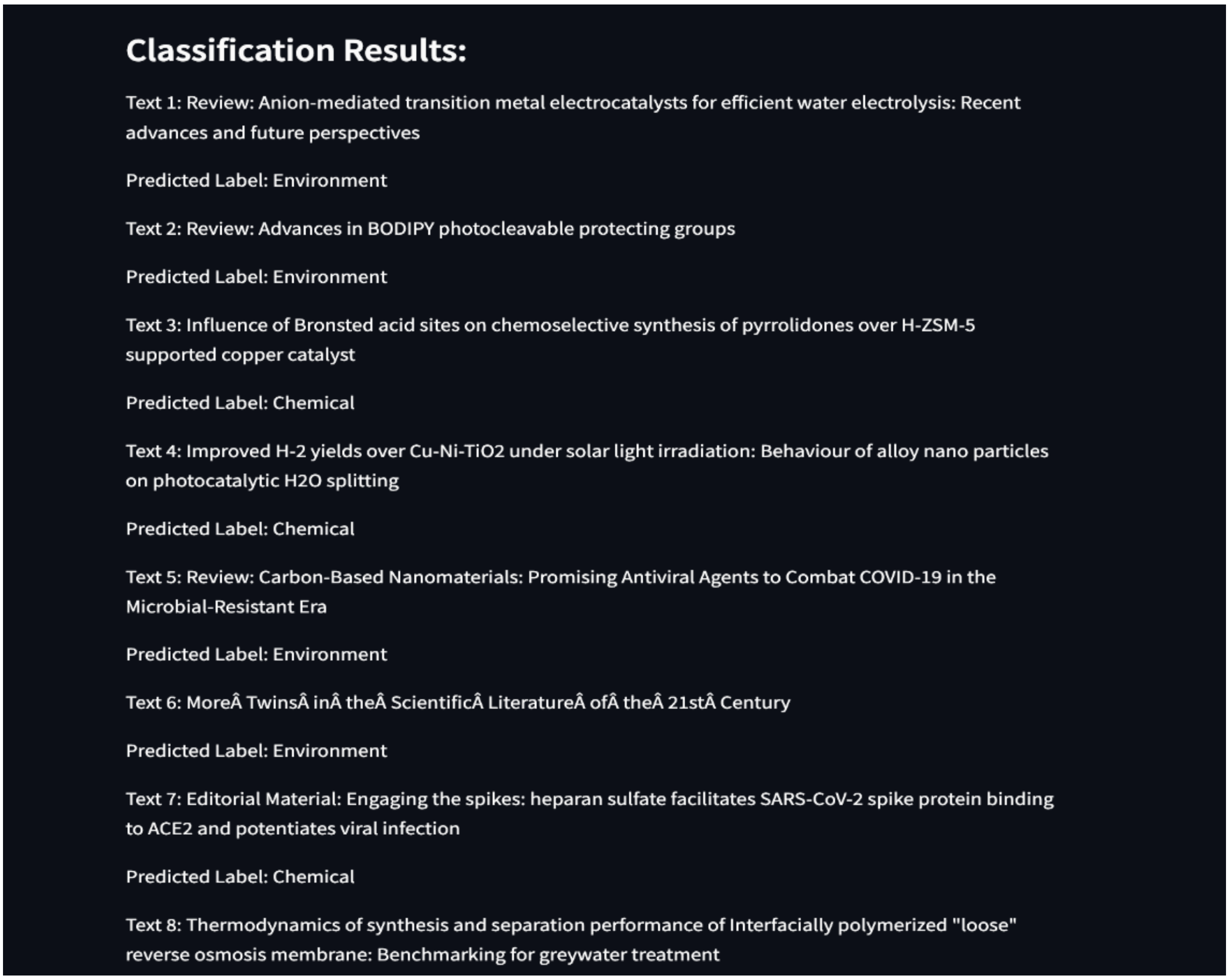
Drag and drop file here
Limit 200MB per file • CSV

Total-Publications-2021-for-CDIS.csv 150.1KB

CSV Data:

	S.No	AUTHOR	TITLE
0	1	BOPPELLA, RAMIREDDY; TAN, JEIWAN; YUN, JUWON; MANORAMA, SUNKARA, V; MOOI	Review: Anion-n
1	2	SINGH, PRAVEEN KUMAR; MAJUMDAR, POULOMI; SINGH, SURYA PRAKASH	Review: Advance
2	3	BOOSA, VENU; VARIMALLA, SHIRISHA; DUMPALAPALLY, MAHESH; GUTTA, NARESH; VE	Influence of Bro
3	4	KUMAR, M. KOTESH; NARESH, G; KUMAR, V. VIJAY; VASISTA, B. SAI; SASIKUMAR, B; VEN	Improved H-2 yi

Fig 5: Web Interface using Streamlit



Classification Results:

Text 1: Review: Anion-mediated transition metal electrocatalysts for efficient water electrolysis: Recent advances and future perspectives

Predicted Label: Environment

Text 2: Review: Advances in BODIPY photocleavable protecting groups

Predicted Label: Environment

Text 3: Influence of Bronsted acid sites on chemoselective synthesis of pyrrolidones over H-ZSM-5 supported copper catalyst

Predicted Label: Chemical

Text 4: Improved H-2 yields over Cu-Ni-TiO2 under solar light irradiation: Behaviour of alloy nano particles on photocatalytic H2O splitting

Predicted Label: Chemical

Text 5: Review: Carbon-Based Nanomaterials: Promising Antiviral Agents to Combat COVID-19 in the Microbial-Resistant Era

Predicted Label: Environment

Text 6: MoreÂ TwinsÂ inÂ theÂ ScientificÂ LiteratureÂ ofÂ theÂ 21stÂ Century

Predicted Label: Environment

Text 7: Editorial Material: Engaging the spikes: heparan sulfate facilitates SARS-CoV-2 spike protein binding to ACE2 and potentiates viral infection

Predicted Label: Chemical

Text 8: Thermodynamics of synthesis and separation performance of Interfacially polymerized "loose" reverse osmosis membrane: Benchmarking for greywater treatment

Fig 6:Classification Results

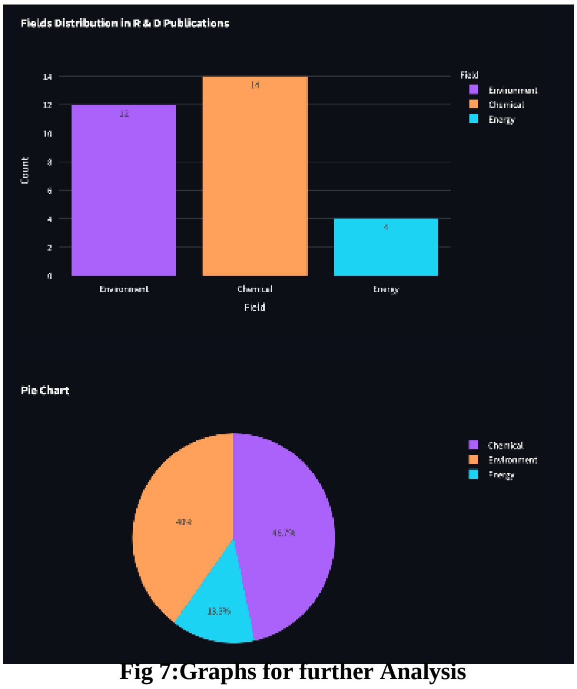
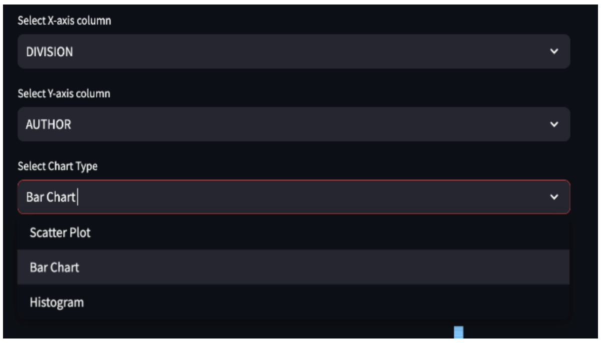


Fig 7:Graphs for further Analysis



Select X-axis column

DIVISION

Select Y-axis column

AUTHOR

Select Chart Type

Bar Chart

Fig 8:Customisable Graph tools for Deep Analysis

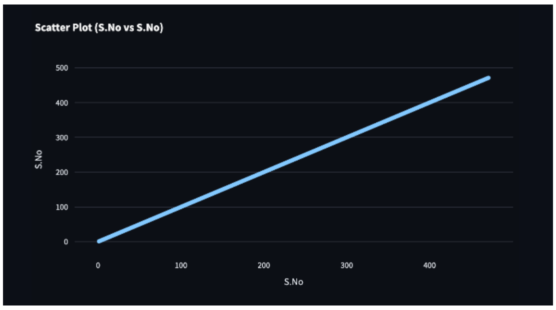


Fig 9:Advanced Visualization

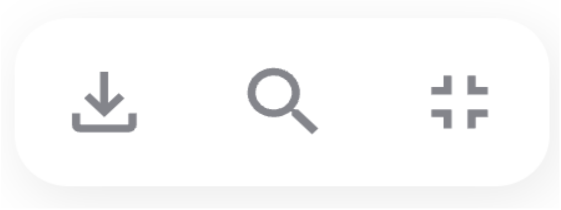


Fig 10:Sort, Search and Download options

7. CONCLUSION

In this study, a supervised approach to zero-shot learning for field classification of texts, leveraging file data to enhance text categorization is explored. By integrating robust language models and strategically utilizing contextual embedding, the potential to accurately classify texts into predefined fields without requiring extensive labeled datasets is demonstrated. This methodology not only reduces the dependency on labor-intensive data labeling but also adapts flexibly to a wide range of classification tasks, highlighting its versatility and scalability.

Our experiments showed that models such as RoBERTa and GPT-3, when applied in a zero-shot learning context, can achieve competitive accuracy across different text categorization tasks. Among these models, RoBERTa consistently outperformed others in terms of precision and recall, underscoring its robustness in understanding and categorizing textual data. The incorporation of positional encodings and multi-head attention mechanisms played a critical role in achieving these results, enabling the models to capture intricate semantic relationships within the texts.

In conclusion, our supervised approach to zero-shot learning for field classification presents a promising solution to the challenges of text categorization. By leveraging advanced language models and file data, one can achieve high accuracy and adaptability, with RoBERTa standing out as the most effective model. This paves the way for more intelligent and automated text processing systems, demonstrating that zero-shot learning is not only viable but also advantageous over traditional methods.

8. FUTURE WORK

The following activities are proposed to enhance our application to be more efficient.

- The application would be accessing not only the titles of publications, but also can be further classified based on the journals they have been published into. This would improve the efficiency of our application as it would be easier to analyze each of the department’s weightage in the renowned journals.
- A mobile application can be built based on the presented work, which would be easier to use and accessible to everyone.
- The present work can be extended to other R & D and academic institutions with necessary customization of relevant data fields.

9. ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to Dr. D. Srinivas Reddy, Director of CSIR-Indian Institute of Chemical Technology (CSIR-IICT) for permitting me to conduct my dissertation work at this esteemed institution**[IICT/Pubs./2024/207]**. Special thanks to Mr.Anil Kumar Pulipaka, Principal Scientist at CSIR-IICT for his invaluable guidance and support throughout this work. I am also grateful to Dr. J. Prasanna Kumar, Head of Department at MVSR Engineering College for allowing me to carry out this work at CSIR-IICT. Additionally, I extend my appreciation to Dr. L. Giribabu, Chairman of Academic Affairs Unit(AAU), CSIR-IICT for his encouragement. Finally, I would like to thank all my family and friends for their unwavering support.

10. REFERENCES

[1] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems, 33*, 1877-1901.

[2] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research, 21*, 1-67.

[3] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

[4] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1*, 4171-4186.

[5] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2019). XLNet Generalized Autoregressive PreTraining for Language Understanding. *Advances in Neural Information Processing Systems, 32*, 5753-5763.

[6] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., &Soricut, R. (2020). ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *International Conference on Learning Representations*.

[7] Radford, A., et al. (2021). "Learning Transferable Visual Models From Natural Language Supervision." arXiv preprint arXiv:2103.00020