A large-scale and high-quality dataset for analyzing research contributions in NLP/ML/AI

Manasa Cherukupally
Department of Information Science
University of North Texas
Frisco, United States of America
ManasaCherukupally@my.unt.edu

Shyam Sundar Domakonda
Department of Information Science
University of North Texas
Denton, United States of America
ShyamSundarDomakonda@my.unt.edu

Yamini Ravala
Department of Information Science
University of North Texas
Denton, United States of America
yaminiravala@my.unt.edu

Sowmya Ushake
Department of Information Science
University of North Texas
Denton, United States of America
SowmyaUshake@my.unt.edu

Soumya Nanditha Chadalavada
Department of Information Science
University of North Texas
Denton, United States of America
SoumyaNandithaChadalavada@my.unt.edu

Sujit Murahari Giridharan
Department of Information Science
University of North Texas
Denton, United States of America
sujitmuraharigiridharan@my.unt.edu

***Abstract*– In this current world research in the areas of Natural Language Processing, Artificial Intelligence, and Machine learning is rapidly increasing. The trends in these technologies are rapidly changing. It is very important to analyze the research papers to understand the trend. In the process of research analysis, the primary focus is on keyword extractions, citations, and structural analyses. But our study focused on the research contributions which are areas less explored. These contributions have novel features and outcomes integral to understanding and advancing the respective field, providing significant insights into the methodology and the research approaches of the paper. Analyzing the contributions will help in a better understanding of the paper. The research project aims to create a high-quality dataset for analyzing research contributions in the fields of Natural Language Processing (NLP), Machine Learning (ML), and Artificial Intelligence (AI). which are categorized into six major types of research contributions. The research methodology involves data collection through web scraping, contribution sentence extraction using a rule-based approach, data cleaning, exploratory data analysis, and classification using models like SciBERT and GPT-2. The paper provides an overview of the entire process, literature review, and detailed results of experiments and analyses conducted.**

GitHub link-

1. **Project folder-** https://github.com/ManasaCherukupally1/Manasa_INFO5731_Spring2023/tree/main/Group3-FinalProject
2. **EDA and Prediction using SciBERT and GPT-2** https://github.com/ManasaCherukupally1/Manasa_INFO5731_Spring2023/blob/main/Group3-FinalProject/SciBERT_GPT2_final.ipynb

***Keywords*– Research Contributions, Dataset, Text classification, Natural Language Processing, Artificial Intelligence, SciBERT, Generative Pretrained Transformers (GPT), Topic Modeling, BERTopic.**

I. INTRODUCTION

Natural Language Processing (NLP) is a vast domain that deals with text processing, classification, and Analysis. Analyzing the research papers helps in understanding the current research practices and the ongoing trends in research methodologies. This analysis is done through various techniques making the sub-part of the research paper like abstract, keywords, and conclusions as a key element for analysis.

A. Background

The whole process of this data extraction from conferences delves into the challenges and intricacies associated with text classification, particularly within the domains of Natural Language Processing (NLP), Machine Learning (ML), and Artificial Intelligence (AI). Prior evaluations in this field primarily focused on keyword frequency, citation counts, and structural analysis. However, this study proposes a unique approach by emphasizing the incorporation of novel knowledge, methodologies, technologies, or insights that were previously unavailable or inadequately explored. The resulting contributions aim to add significant value to the academic community and potentially influence future research, steering the trajectory of the field.

B. Challenges in Contribution Sentences Collection

The primary challenges encountered in this research pertain to the vast amounts of data present in conferences, making data cleaning and web scraping arduous tasks. The sheer volume of input data complicates the extraction of essential details such as authors, their contributions, and the year of publication. Moreover, some conferences impose restrictions on IP addresses, hindering data access and leading to runtime errors in the code.

C. Research Objectives

The primary motive of this project is to build a high-quality research contribution dataset that is categorized into six different categories. This dataset can be helpful in the automated summarization of the research papers based on their contribution sentences. The categories that were used in this project are

1. Algorithms/Methods Construction or Optimization
2. Performance Evaluation
3. Model Construction or Optimization
4. Data Creation or Resources
5. Applications (repeated for emphasis)
6. Theory Proposal

To achieve this research objective, certain steps are followed across the project. Starting with collecting the data from the conferences. Followed by extracting the contribution sentences. Once the contributions are extracted and ready the model is trained with high quality labeled dataset provided by Chen et. al. (2022). The trained model is evaluated for performance. Finally, the extracted dataset is classified into six categories through classification using the Hugging Faces Transformers model.

This comprehensive approach combines rule-based extraction, state-of-the-art model training with SCI-BERT, and evaluation using GPT-2 to create a sophisticated framework for classifying contributions in the legal text classification domain. The resulting categorized dataset is expected to provide valuable insights for researchers and practitioners in NLP, ML, and AI, shaping the future trajectory of research in this dynamic field.

D. Overview of Paper

This paper explains the end-to-end process of classifying the research contributions into six major categories. The Literature review related to the project is explained in section 2. Section 3 describes the data collection process. Section 4 explains the methodology including the data cleaning, Exploratory Data Analysis, and Analysis techniques used. Section 5 talks about the experiments conducted and the parameter tuning performed. Section 6 discusses the obtained results and other discussions. Followed by the Conclusion and Limitations of the project in Section 7.

II. LITERATURE REVIEW

A. Extraction of Research Contributions

Constructing a large-scale and high-quality dataset is a tedious task and it requires a lot of research and studying other papers where (Sonal et al.. 2011) research contributions are extracted by categorizing articles into Focus, Technique, and Domain and (Bolin Hua et al..2021) focused on extracting sentences of originality from Conclusion section. (S. Otani et al..2014) extracted key expressions from article abstracts and (Aurelie et al..2011) extracted data deposition statements from the literature. However, focusing on only one or two aspects of a research paper to extract contribution sentences could result in inconsistencies and information loss. The rule-based method (Bolin Hua et al..2021) imposes key rules or conditions to extract contributions and find key features. This method seemed feasible to extract contribution sentences from articles. (Li.y..et al..2022) constructed a Chinese scientific literature dataset that includes information on academic papers such as abstracts, titles, and keywords and used a manual annotation scheme to categorize them. We followed the fine grain annotation scheme (Chen et al..2022) to categorize the extracted contribution sentences.

B. Identification and Analysis of Contribution Sentences

The importance of extracting both subjective and objective features when using rule-based methods is emphasized by (Yin k et al..2017). This approach improved the recall of extraction and reinforced the extraction of contextual sentences. The extracted tweets on climate change were analyzed by pre-processing the opinions (Samson et al..2023) which included stop words removal, URL tags removal, etc.. These techniques proved helpful in eliminating unnecessary factors and left important words and topics that can be further evaluated. Topic models have been used to study the popularity of communities (Griffiths and Steyvers, 2004), the history of ideas (Hall et al., 2008), and the scholarly impact of papers (Gerrish and Blei, 2010). Topic modeling techniques such as LDA and BERT (Samson et al..2023) narrow down the contextual relation categorize the text into different clusters and determine their importance. These techniques map words together and determine their frequency. Cloud models (Wang et al..2023) prove to be efficient while determining the most repeated topics and highlighting them.

C. Machine Learning for Classification of Contributions

Numerous machine learning techniques are being used for the classification of text such as reviews, research contributions, tweets, etc... (Wang et al., 2023) used a deep learning approach, a BERT model, and a back cloud generator for classification and had good results. (Auer et al.. 2021) proposed a SemEval-2021 Shared Task NLP Contribution Graph that would automatically structure the contributions together and produce a knowledge graph but it has different annotation schemes at the sentence level, and phrase level. CiteOpinion (Le.x et al..) is a tool that focuses on identifying citation sentences and analyzing them for academic contributions using DL. Li et al. (2020) compared the multiple ML and DL models for text classification. Among these, the transformer-based methods (i.e., ELMo, GPT, BERT), which apply unsupervised methods to mine semantic knowledge automatically and then construct pre-training targets to support semantic understanding, have been widely used and proven effective. (Chen et al..2022) included the SciBERT model along with other ML models and it provided greater results when compared to other models. The BERT model proved to be more efficient in classification tasks. Classification accuracy is improved (Shaikh et al.. 2021) by handling imbalances in datasets by using

conventional ML, LSTM, and GPT-2 models. GPT 2 showed high accuracy.

GPT-2 is efficient at training samples for improving classification accuracy.(Edwards et al..2021). (E.T.R et al..2021) used the GPT-2 model to train samples on scientific Portuguese text and classify them and it outperformed generic models. The large transformer models have been trained on large data samples and build deep contextual relations between the text which makes it a good choice to train samples on for classification tasks.

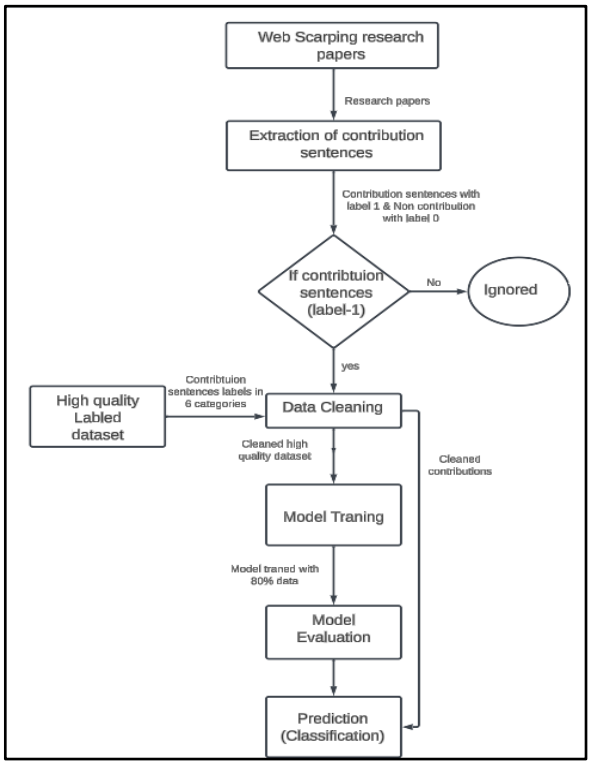


fig. 1. Flow chart of the research process

III. Data Collection and Exploratory Data Analysis

A. Web Scraping to collect research papers

We have used web scraping techniques to collect the research papers from different conferences. The conferences that are webscraped for this project include JMLR, CoNLL, and EMNLP which has around 9500+ research papers. We have tried to scrape from other websites as well but due to access restrictions, the list of conferences was limited to these three. Python package Beautifulsoup is used to extract the research papers. The "requests" library is used to make HTTP requests to the conference URL. Research papers from the past 10 years have been collected. The scraped research papers are arranged according to the year published. "fig. 2" shows the sample of scraped research papers.

B. Extracting Contribution Sentences

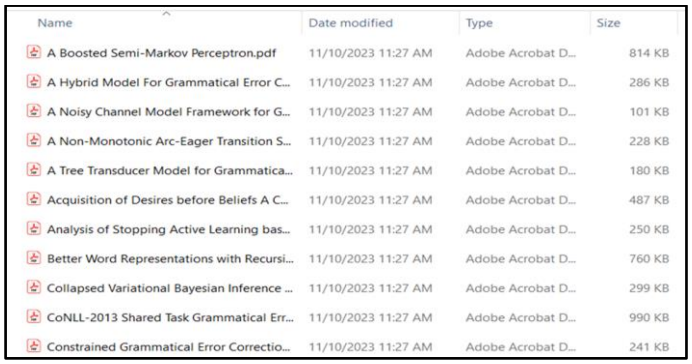
For extracting the contribution sentences from the scraped research papers Rule-Based approach is used. As part of this approach, "Pronoun followed by Verb" is used. Pronouns like ['we', 'this paper', 'in this paper', 'i', 'the authors', 'our', 'the study', 'the research'] and Verbs like ['propose', 'contribute', 'introduce', 'describe', 'present', 'report', 'discuss', 'employ', 'study', 'explore', 'aim', 'hypothesize', 'argue', 're-use', 'evaluate', 'compare', 'baseline', 'result', 'approach', 'method', 'extend', 'ensemble', 'transfer learning', 'architecture', 'neural method', 'translation systems', 'system development', 'domain adaptation', 'train', 'experiment'] are matched to extract contribution sentences. Extracted sentences are labeled as "1" for contribution sentences and "0" for noncontribution sentences for which we got 385224 sentences in total. Fig 2 shows the sample sentences extracted.

C. Data Cleaning

From the entire dataset, only contribution sentences are selected which are of count 192616. For further analysis, sentences that have < than 50 words are dropped to improve the quality. We considered only the first 3000 contribution sentences since the data is very huge. Apart from that several data-cleaning techniques are applied on the extracted contribution sentences. These techniques include

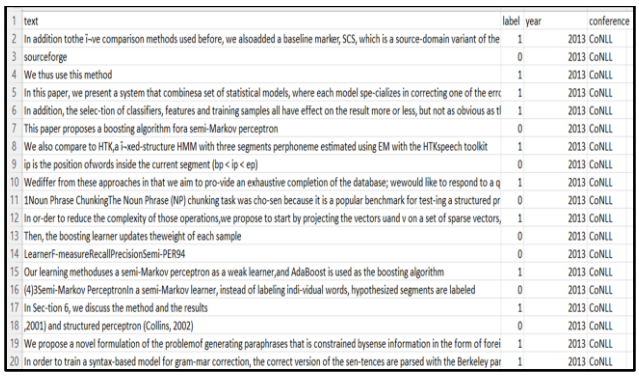
1. Remove URLs, HTML, Tags, and Mentions
2. Remove punctuations and ASCII Characters
3. Lemmatization
4. Remove Stop Words

The cleaned dataset sample is presented in Fig. 4.



Name	Date modified	Type	Size
A Boosted Semi-Markov Perceptron.pdf	11/10/2023 11:27 AM	Adobe Acrobat D...	814 KB
A Hybrid Model For Grammatical Error C...	11/10/2023 11:27 AM	Adobe Acrobat D...	286 KB
A Noisy Channel Model Framework for G...	11/10/2023 11:27 AM	Adobe Acrobat D...	101 KB
A Non-Monotonic Arc-Eager Transition S...	11/10/2023 11:27 AM	Adobe Acrobat D...	228 KB
A Tree Transducer Model for Grammatica...	11/10/2023 11:27 AM	Adobe Acrobat D...	180 KB
Acquisition of Desires before Beliefs A C...	11/10/2023 11:27 AM	Adobe Acrobat D...	487 KB
Analysis of Stopping Active Learning bas...	11/10/2023 11:27 AM	Adobe Acrobat D...	250 KB
Better Word Representations with Recursi...	11/10/2023 11:27 AM	Adobe Acrobat D...	760 KB
Collapsed Variational Bayesian Inference ...	11/10/2023 11:27 AM	Adobe Acrobat D...	299 KB
CoNLL-2013 Shared Task Grammatical Err...	11/10/2023 11:27 AM	Adobe Acrobat D...	990 KB
Constrained Grammatical Error Correctio...	11/10/2023 11:27 AM	Adobe Acrobat D...	241 KB

fig. 2 Sample of scraped research paper



	text	label	year	conference
2	In addition tothe i–ve comparison methods used before, we alsoadded a baseline marker, SCS, which is a source-domain variant of the	1	2013	CoNLL
3	sourceforge	0	2013	CoNLL
4	We thus use this method	1	2013	CoNLL
5	In this paper, we present a system that combinesa set of statistical models, where each model spe-cializes in correcting one of the errc	1	2013	CoNLL
6	In addition, the selec-tion of classifiers, features and training samples all have effect on the result more or less, but not as obvious as tl	1	2013	CoNLL
7	This paper proposes a boosting algorithm fora semi-Markov perceptron	0	2013	CoNLL
8	We also compare to HTK,a i–xed-structure HMM with three segments perphoneme estimated using EM with the HTKspeech toolkit	1	2013	CoNLL
9	ip is the position ofwords inside the current segment (bp < ip < ep)	0	2013	CoNLL
10	Wediffer from these approaches in that we aim to pro-vide an exhaustive completion of the database; wewould like to respond to a q	1	2013	CoNLL
11	1Noun Phrase ChunkingThe Noun Phrase (NP) chunking task was cho-sen because it is a popular benchmark for test-ing a structured pr	0	2013	CoNLL
12	In or-der to reduce the complexity of those operations,we propose to start by projecting the vectors uand v on a set of sparse vectors,	1	2013	CoNLL
13	Then, the boosting learner updates theweight of each sample	0	2013	CoNLL
14	LearnerF-measureRecallPrecisionSemi-PER94	0	2013	CoNLL
15	Our learning methoduses a semi-Markov perceptron as a weak learner,and AdaBoost is used as the boosting algorithm	1	2013	CoNLL
16	(4)3Semi-Markov PerceptronIn a semi-Markov learner, instead of labeling indi-vidual words, hypothesized segments are labeled	0	2013	CoNLL
17	In Sec-tion 6, we discuss the method and the results	1	2013	CoNLL
18	,2001) and structured perceptron (Collins, 2002)	0	2013	CoNLL
19	We propose a novel formulation of the problemof generating paraphrases that is constrained bysense information in the form of forei	1	2013	CoNLL
20	In order to train a syntax-based model for gram-mar correction, the correct version of the sen-tences are parsed with the Berkeley par	1	2013	CoNLL

fig. 3 Sample dataset of contribution and non-contribution sentences

	Text	label	year	conference
0	present approach based neu ral network model a...	1.0	2013.0	EMNLP
1	contribution present tmi bootstrap ping system...	1.0	2013.0	EMNLP
2	evaluate pro posed summarization approach tac ...	1.0	2013.0	EMNLP
3	use pagerankbased approach proposed yang	1.0	2013.0	EMNLP
4	compared approach two classic simple method fe...	1.0	2013.0	EMNLP
...	...	...	...	...
2995	train distilbert hate speech classification mo...	1.0	2022.0	CoNLL
2996	presented new approach unsupervised cognate id...	1.0	2022.0	CoNLL
2997	researchquestion level structural informationi...	1.0	2022.0	CoNLL
2998	proposed amethod estimating syntactic predicta...	1.0	2022.0	CoNLL
2999	wemake implementation publicly available8in ho...	1.0	2022.0	CoNLL

fig. 4. Cleaned Contribution sentences

D. *Exploratory Data Analysis*

Several data analysis techniques are used for initial analysis. As part of this step, Word count, Average word length in each sentence, and Sentence length for each conference are analyzed. The Box-plot and KDE show the results of this analysis. The top ten Unigrams, Bigrams, and Trigrams are extracted from the data and it found that "model", "method" and "approach" are top unigrams. "paper proposes", "data set", and neural networks are top Bigrams. Word clouds with the most frequent words for each conference were extracted. The top 10 words with the highest TF-IDF values are extracted for each sentence. POS-Tagging counts are calculated for each conference.

Top 10 1-grams:		Top 10 2-grams:		Top 10 3-grams:	
model	612	paper propose	49	draw draw draw	9
method	546	data set	40	fixed fixed fixed	6
approach	407	neural network	36	random random random	6
result	373	report result	36	also compare result	5
section	281	test set	33	applied approach generate	5
present	272	paper present	28	approach generate propbanks	5
data	238	train model	26	crosslingual dependency parsing	5
propose	222	language model	24	future would like	5
also	214	training data	24	generate propbanks language	5
set	190	baseline model	21	language group akbik	5
				dtype: int64	

fig. 5. Unigrams, Bigrams and Trigrams

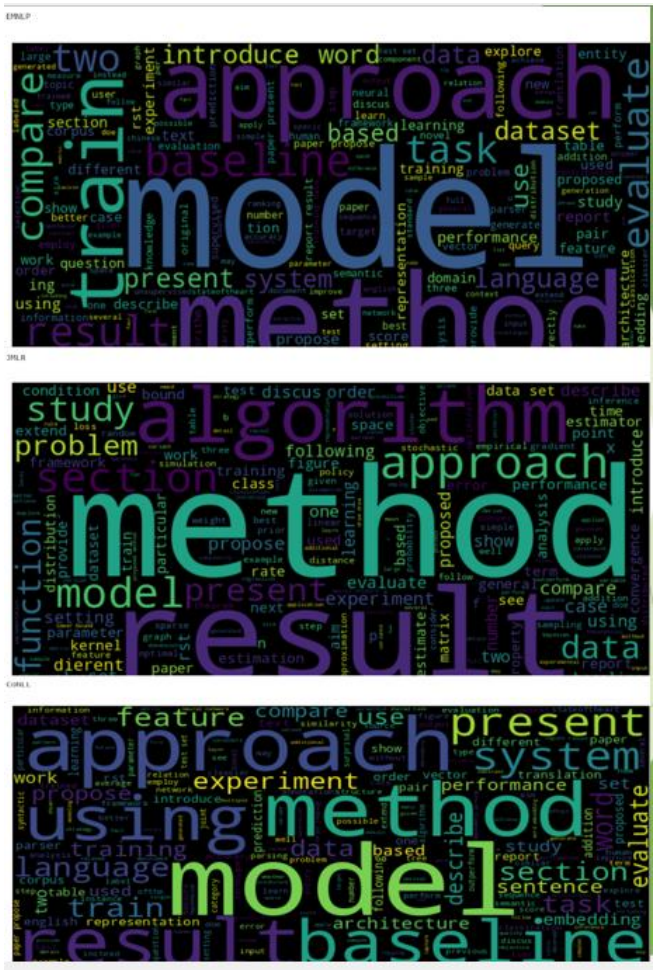


fig. 6. Word cloud showing top frequency words for each conference in the order EMNLP,JMLR, CoNLL

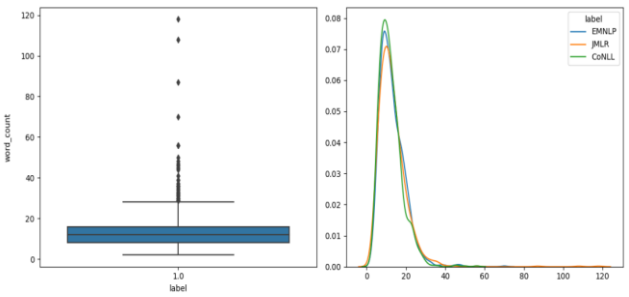


fig. 7.1. Boxplot and KDE of Wordcount

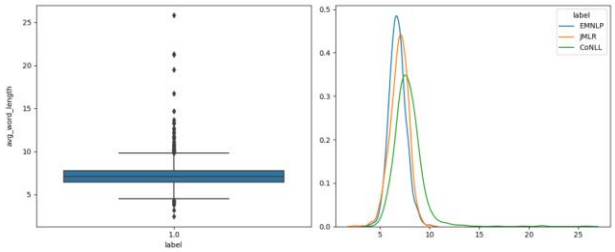


fig. 7.2. Boxplot and KDE of Average word length

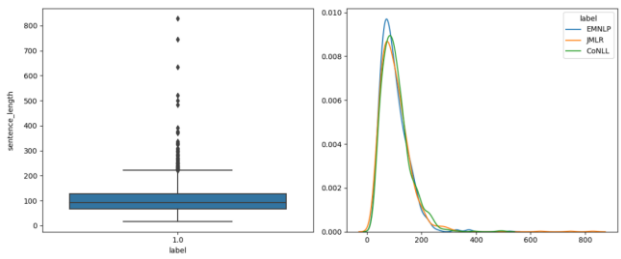


fig. 7.3. Boxplot and KDE of Sentence length

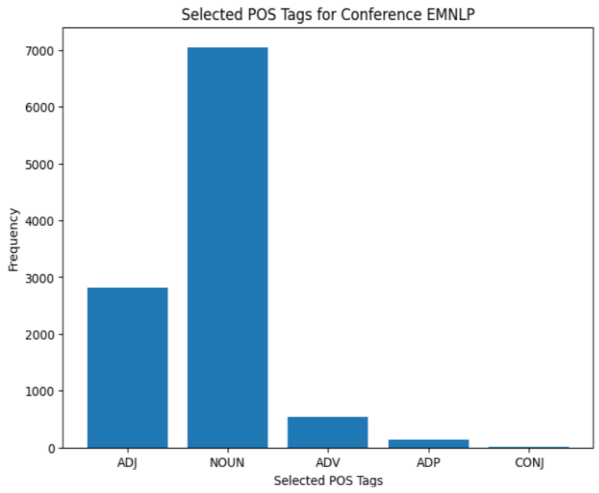


fig. 8.1 POS-Tagging of words in EMNLP

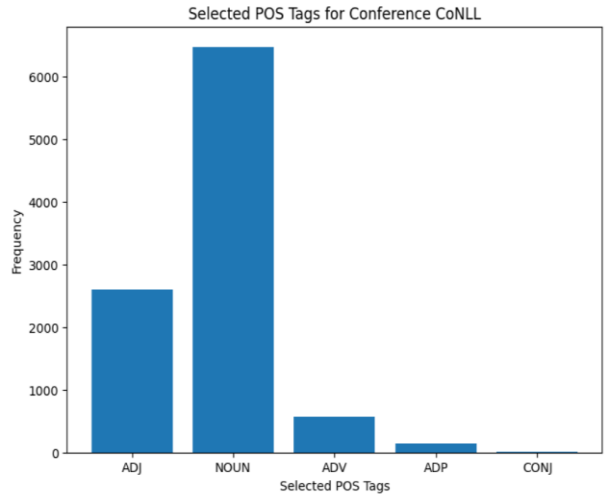


fig. 8.2 POS-Tagging of words in CoNLL

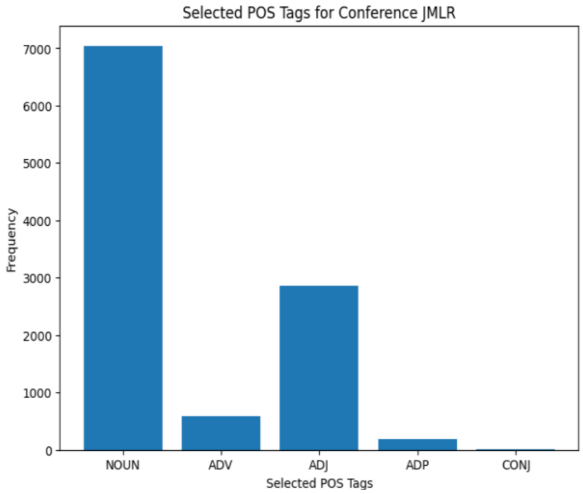


fig. 8.3 POS-Tagging of words in JMLR

E. Topic Modeling

Two Topic modeling techniques, LDA and BERTopic are used to extract the topics. LDA has given 20 topics whereas BERTopic has given 42 topics. BERTopic has given better topics when compared to LDA. The top 20

topics given by BERTopics are given in Fig. 10. The top 3 topics are "model_method_approach_result", "section_discuss_work_describe" and "gradient_method_problem_solution". Topic 1 is closer to most topics.

	Topic	Count	Name
0	-1	1599	-1_model_method_approach_result
1	0	184	0_section_discus_work_describe
2	1	88	1_gradient_method_problem_solution
3	2	72	2_neural_network_architecture_rnn
4	3	71	3_embeddings_word_embedding_vector
5	4	66	4_report_table_average_result
6	5	52	5_policy_learning_bandit_agent
7	6	44	6_rst_decoding_review_section
8	7	43	7_test_validation_set_training
9	8	42	8_graph_causal_matching_graphbased
10	9	39	9_experiment_section_experimental_example
11	10	37	10_lemma_theorem_proof_bound
12	11	36	11_bayesian_posterior_distribution_prior
13	12	35	12_algorithm_method_fqe_compare
14	13	35	13_fixed_draw_map_random
15	14	34	14_parser_dependency_parsing_tree
16	15	30	15_ranking_candidate_bsus_list
17	16	25	16_translation_appreciation_query_player
18	17	25	17_english_language_bilingual_glish
19	18	24	18_speech_sound_dialogue_acoustic
20	19	23	19_similarity_word_semantic_measure

fig. 9. Top 20 topics out of 42 topicsgiven by BERTopic

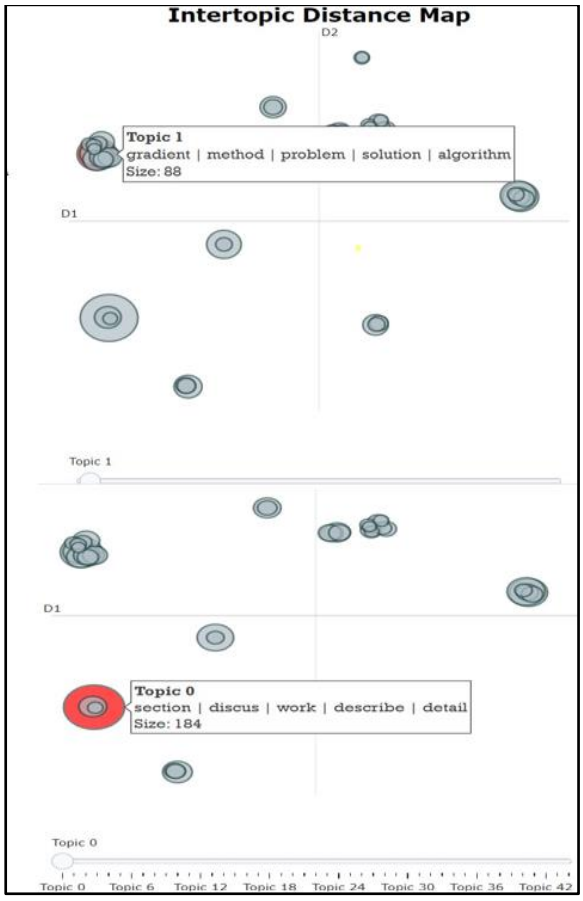


fig. 10.Intertopic distance showing the Topics closer to Topic-1 and Topic-0

IV. Methodology

A. Model Traning & Evalaution

In order to train the model we have used the high quality dataset contributed by the Chen et. al. (2022) available in the https://zenodo.org/records/6284137#.YhkZ7-iZO4Q . This dataset consists of 5025 contribution sentences classified into 6 categories Algorithms/ Methods Construction or Optimization, Applications, Model Construction or Optimization, Data Creation or Resources, Applications, and Theory Proposal. Two classification models were selected to train from the Huggingface Transformers library. Huggiface is a company known for open-source tools and libraries for NLP tasks.

a. SciBERT Model for Text Classification

SciBERT is one of the most popular models for text classification. The reason for selecting SciBERT is because it is trained on scientific data and works well for processing and classification of scientific text. This pre-trained model is used for our text classification. The high-quality dataset selected is split into training and testing data in the ratio of 8:2. The SciBERT model is fine-tuned with 80% of the high-quality dataset selected. This fine-tuned model is tested and evaluated on 20% of the data.

b. Generative Pre-trained Transformers-GPT-2

OpenAI's GPT-2 model is the trending model for text classification and text processing. GPT-2 model is a Generative Pre-trained model trained on billions of textual data. Hence GPT-2 is selected for training the model. From the Transformers library, GPT2ForSequenceClassification is used for training with the selected dataset. 10 epochs with 32 batch size is given while training the model. This pre-trained model is finetuned with 80% of the high-quality dataset. The other 20% is used for testing and evaluating the Model

Both the models trained are evaluated using Classification metrics like accuracy, Precision, and Recall. A detailed explanation of the Python packages used and the parameters applied are discussed in the experiments section.

B. Predicting the Extracted Contribution Sentences

The dataset with the extracted contribution sentences which are cleaned and ready to predict are applied to both of the models to predict the labels and categorize them into 6 categories. This dataset after prediction holds the contribution sentences that are classified into 6 categories and can be used for analyzing the research papers based on their respective contributions making it a high-quality dataset.

```
GPT2ForSequenceClassification(
  (transformer): GPT2Model(
    (wte): Embedding(50257, 768)
    (wpe): Embedding(1024, 768)
    (drop): Dropout(p=0.1, inplace=False)
    (h): ModuleList(
      (0-11): 12 x GPT2Block(
        (ln_1): LayerNorm((768,), eps=1e-05, elementwise_affine=True)
        (attn): GPT2Attention(
          (c_attn): Conv1D()
          (c_proj): Conv1D()
          (attn_dropout): Dropout(p=0.1, inplace=False)
          (resid_dropout): Dropout(p=0.1, inplace=False)
        )
        (ln_2): LayerNorm((768,), eps=1e-05, elementwise_affine=True)
        (mlp): GPT2MLP(
          (c_fc): Conv1D()
          (c_proj): Conv1D()
          (act): NewGELUActivation()
          (dropout): Dropout(p=0.1, inplace=False)
        )
      )
    )
    (ln_f): LayerNorm((768,), eps=1e-05, elementwise_affine=True)
  )
  (score): Linear(in_features=768, out_features=6, bias=False)
)
```

fig. 10. GPT-2 model built for classification

V. Experiment

A. Fine-Tuning SciBERT for Text Classification: An In-depth Exploration and Parameter Analysis.

In this experiment, we embarked on the task of fine-tuning the SciBERT model for a specific text classification objective. The process commenced with the loading of essential SciBERT tokenizer and model components through the Hugging Face Transformers library. The dataset, comprising textual information, underwent tokenization and padding to a standardized maximum length of 512 tokens. Following this, categorical labels were encoded into a numerical format using the LabelEncoder. PyTorch DataLoader objects were then instantiated to facilitate efficient batch processing during both training and testing phases.

The fine-tuning process involved the utilization of the AdamW optimizer with a learning rate of 2e-5 across three training epochs. To bolster the model's stability during training, a linear learning rate scheduler was implemented, accompanied by a gradient clipping technique to address potential issues related to exploding gradients. The training loop diligently monitored the average training loss after each epoch. Validation on a separate dataset was conducted after each epoch, facilitating the identification and retention of the best-performing model based on the highest validation accuracy.

For subsequent evaluation on the test set, the trained model was employed to generate predictions, subsequently compared with the ground truth labels. The resulting accuracy on the test set was approximately 58.11%, providing valuable insights into the model's proficiency in handling the specific text classification task under consideration.

Key parameters critical to the experiment included a maximum token length of 512, an AdamW optimizer with a learning rate of 2e-5, and a batch size of 8. The training regimen encompassed three epochs, supported by a linear learning rate scheduler. The incorporation of gradient clipping during training aimed to mitigate potential challenges associated with gradient explosions.

B. *Optimizing Text Classification with GPT-2: Parameter Tuning and Performance Analysis*

In this experiment, the focus was on fine-tuning the GPT-2 model for a text classification task, specifically aimed at achieving enhanced performance through careful parameter selection. The dataset was preprocessed, and a subset was chosen for model training, while the remaining samples were reserved for validation. The key parameters included a maximum sequence length of text, set to 'None' for variable lengths, a batch size of 32, and the choice of running the training process for 10 epochs.

A custom dataset class, `DatasetCreator`, was implemented to structure the data for PyTorch, allowing seamless integration with the DataLoader. The GPT2 tokenizer and collator were employed to process and prepare the text inputs for the model. The training loop incorporated the AdamW optimizer with a learning rate of 5e-5, epsilon of 1e-8, and weight decay of 0.01. Additionally, a linear learning rate scheduler with warm-up steps was introduced to optimize training stability.

Throughout the training process, the model was fine-tuned using the labeled dataset, and performance metrics such as training accuracy and loss were tracked. The validation set was leveraged to assess the model's generalization capability, with accuracy and loss recorded at each epoch.

The experiment concluded after 9 epochs, and the validation accuracy achieved was approximately 0.56. This result provides insights into the model's effectiveness in classifying text for the specific task under consideration. The emphasis on parameter tuning and performance monitoring contributes to a comprehensive understanding of the GPT-2 model's behavior in the context of text classification without delving into specific code details or data-splitting strategies.

VI. Results and Discussions

A. *SciBERT Model*

The model has an overall accuracy of 58.11%. From Algorithms/ Methods Construction or Optimization to Theory Proposal, each category is rigorously examined for precision, recall, and the harmonizing f1-score.

Classification Report:

	precision	recall	f1-score	support
Algorithms/ Methods Construction or Optimization	0.56	0.66	0.61	243
Applications	0.48	0.26	0.34	46
Dataset Creation or Resources	0.58	0.47	0.52	113
Model Construction or Optimization	0.56	0.58	0.57	197
Performance Evaluation	0.52	0.46	0.49	133
Theory Proposal	0.65	0.67	0.66	273
accuracy			0.58	1005
macro avg	0.56	0.52	0.53	1005
weighted avg	0.58	0.58	0.58	1005

fig. 10. Classification Report for SciBERT

Theory Proposal has the greatest f1-score, indicating a noteworthy balance in both the accuracy of the model's predictions (precision) and its sensitivity to the actual data (recall). The number of data points (support) in each category provides information on the distribution of the dataset, which is important for understanding the model's metrics in the context of data abundance or scarcity.

In the fig below, the Algorithms/ Methods Construction or Optimization has the most prominent category, with 1,036 instances, highlighting a large representation of works focusing on the development and advancement of computational techniques within the dataset. Applications is the least represented category, with only 90 instances, indicating a comparatively lower emphasis on applied research. This distribution represents the corpus' underlying topic focus, providing a quantitative view on current research trends.

```
Algorithms/ Methods Construction or Optimization    1036
Theory Proposal                                       593
Performance Evaluation                                567
Model Construction or Optimization                    431
Dataset Creation or Resources                         283
Applications                                           90
Name: Predicted_Label, dtype: int64
```

fig. 10. Category Count using SciBERT

B. *ChatGPT 2 Model*

The ChatGPT 2 model received an accuracy of 56%. The most common was Algorithms/ Methods Construction or Optimization with 962 instances, indicating a substantial volume of work in establishing new computational techniques. Theory Proposal and Performance Evaluation come next, with 731 and 669 counts, indicating active areas of theoretical and practical breakthroughs, respectively. The lower frequency in the Dataset Creation or Resources and Applications categories, with 124 and 71 counts,

respectively, indicates that these areas are underrepresented in the dataset.

```
Algorithms/ Methods Construction or Optimization    962
Theory Proposal                                      731
Performance Evaluation                               669
Model Construction or Optimization                   443
Dataset Creation or Resources                        124
Applications                                          71
Name: target, dtype: int64
```

fig. 10. Category Count for ChatGPT 2

SciBERT was marginally more accurate than GPT-2, completing 58% of tasks correctly compared to 56% for GPT-2.

VII. Conclusion and Future Work

In this paper, we constructed a large-scale and high-quality dataset of research contributions from research papers related to NLP, ML, and AI. These research contributions are collected from research papers of conferences CoNLL, EmNLP, and JMLR by using Rule based approach. After the preprocessing, the retrieved contribution sentences were classified into six categories namely Algorithms/Methods Construction or Optimization, Performance Evaluation, Model Construction or Optimization, Data Creation or Resources, Applications (repeated for emphasis), and Theory Proposal. This has been done by training and testing the research contributions using SciBERT and GPT-2 models, demonstrating the usage of cutting-edge machine learning methods. SciBERT delivered an accuracy of 58% and Chatgpt 2 had an accuracy of 56%. They performed well in domains such as Theory Proposals and Algorithms/Methods Construction. Especially in these technical categories, SciBERT demonstrated exceptionally good recall and precision. The limitations are the poor accuracy of both models for classification, this could be further improved training on more no.of epochs and using a model for identifying contribution sentences from non-contribution sentences.

Future work involves improving the SciBERT and ChatGPT 2 models to achieve higher accuracy and interpretability with the possible use of Prompt Engineering.It may also involve investigating newer models that have been developed since the project's conclusion.

VIII Author Contributions

1. *Manasa Cherukupally*- Involved in coding(Collection of research papers,extracting contributions, Data cleaning and model building GPT-2)Involved in making Presentation and Report
2. *Yamini Ravala*- Been part of extracting the data and data cleaning and also did model training. Contributed to making presentations and report.
3. *Shyam Sundar Domakonda*- Engaged in extracting data from conference papers, performing data preprocessing, and building SciBERT. Conducted hyperparameter tuning to enhance the performance of SciBERT. Involved to the creation of the presentation and report.
4. *Sujit Murahari Giridharan*- Extracted the data, involved in coding to extract data collection and contributed to making the Presentation and report.
5. *Soumya Nanditha Chadalavada*- Been part of extracting the data and contributed in making the Presentation and Report
6. *Sowmya Ushake*- I retrieved information, participated in coding to gather data, and contributed to the creation of both the presentation and the report.

References

[1] Chen, H., Nguyen, H., & Alghamdi, A. (2022). Constructing a high-quality dataset for automated creation of summaries of fundamental contributions of research articles. *Scientometrics*, *127*(12), 7061-7075.

[2] Wang, Z., Zhang, H., Chen, J., & Chen, H. Measuring the Novelty of Scientific Literature Through Contribution Sentence Analysis Using Deep Learning and Cloud Model. *Available at SSRN 4360535*.

[3] Ma, X., Wang, J., & Zhang, X. (2021, August). YNU-HPCC at SemEval-2021 Task 11: Using a BERT Model to Extract Contributions from NLP Scholarly Articles. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)* (pp. 478-484).

[4] Sonal Gupta and Christopher Manning. 2011. Analyzing the Dynamics of Research by Extracting Key Aspects of Scientific Papers. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1–9, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

[5] Bolin Hua, Youngkug Shin. Extraction of Sentences Describing Originality from Conclusion in Academic Papers. In Yi Zhang 0042, Chengzhi Zhang, Philipp Mayr 0001, Arho Suominen, editors, *Proceedings of the 1st Workshop on AI + Informetrics (AII2021) co-located with the iConference 2021, Virtual Event, March 17th, 2021*. Volume 2871 of *CEUR Workshop Proceedings*, pages 58-70, CEUR-WS.org, 2021. [doi]

[6] S. Otani and Y. Tomiura, "Extraction of Key Expressions Indicating the Important Sentence from Article Abstracts," 2014 IIAI 3rd International Conference on Advanced Applied Informatics, Kokura, Japan, 2014, pp. 216-219, doi: 10.1109/IIAI-AAI.2014.53.

[7] Aurélie Névéol, W. John Wilbur, Zhiyong Lu, Extraction of data deposition statements from the literature: a method for automatically tracking research results, Bioinformatics, Volume 27, Issue 23, December 2011, Pages 3306–3312, https://doi.org/10.1093/bioinformatics/btr573.

[8] Li, Y., Zhang, Y., Zhao, Z., Shen, L., Liu, W., Mao, W., & Zhang, H. (2022). CSL: A Large-scale Chinese Scientific Literature Dataset. International Conference on Computational Linguistics.

[9] Chen, H., Nguyen, H., & Alghamdi, A. (2022). Constructing a high-quality dataset for automated creation of summaries of fundamental contributions of research articles. Scientometrics, 127(12), 7061-7075.

[10] Yin Kang, Lina Zhou, RubE: Rule-based methods for extracting product features from online consumer reviews, Information & Management, Volume 54, Issue 2, 2017, Pages 166-176, ISSN 0378-7206, https://doi.org/10.1016/j.im.2016.05.007.

[11] T. L. Griffiths and M. Steyvers. 2004. Finding scientific topics. Proceedings of the National Academy of Sciences.

[12] David Hall, Daniel Jurafsky, and Christopher D. Manning. 2008. Studying the history of ideas using topic models. In Proceedings of the Conference on Empirical Methods in Natural Language Processing.

[13] Sean M. Gerrish and David M. Blei. 2010. A language-based approach to measuring scholarly impact. In Proceedings of the International Conference on Machine Learning.

[14] Samson Ebenezar Uthirapathy, Domnic Sandanam, Topic Modelling and Opinion Analysis On Climate Change Twitter Data Using LDA And BERT Model.,Procedia Computer Science, Volume 218, 2023,Pages 908-917,ISSN 1877-0509,https://doi.org/10.1016/j.procs.2023.01.071.

[15] Wang, Z., Zhang, H., Chen, J., & Chen, H. Measuring the Novelty of Scientific Literature Through Contribution Sentence Analysis Using Deep Learning and Cloud Model. Available at SSRN 4360535.

[16] Auer, S., & Pedersen, T. (2021). SemEval-2021 Task 11:NLPContributionGraph -- Structuring Scholarly NLP Contributions for a Research Knowledge Graph. ArXiv. https://doi.org/10.18653/v1/2021.semeval-1.44

[17] Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., Yu, P. S., & He, L. (2020). A survey on text classifca[1]tion: From shallow to deep learning. arXiv preprint arXiv:2008.00364

[18] Shaikh, S.; Daudpota, S.M.; Imran, A.S.; Kastrati, Z. Towards Improved Classification Accuracy on Highly Imbalanced Text Dataset Using Deep Neural Language Models. Appl. Sci. 2021, 11, 869. https://doi.org/10.3390/app11020869

[19] Edwards, A., Ushio, A., De Ribaupierre, H., & Preece, A. (2021). Guiding Generative Language Models for Data Augmentation in Few-Shot Text Classification. ArXiv. /abs/2111.09064

[20] E. T. R. Schneider, J. V. A. de Souza, Y. B. Gumiel, C. Moro and E. C. Paraiso, "A GPT-2 Language Model for Biomedical Texts in Portuguese," 2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS), Aveiro, Portugal, 2021, pp. 474-479, doi: 10.1109/CBMS52027.2021.00056.

[21]Le,X.,Chu,J.,Deng,S.,Jiao,Q.,Pei,J.,Zhu,L.& Yao,J.(2019).CiteOpinion: Evidence-based Evaluation Tool for Academic Contributions of Research Papers Based on Citing Sentences. Journal of Data and Information Science,4(4) 26-41. https://doi.org/10.2478/jdis-2019-0019