Automated Citation Function Classification and Context Extraction in Astrophysics: Leveraging Paraphrasing and Question Answering

Hariram Veeramani
Department of Electrical and Computer Engineering, UCLA, USA
hariram@ucla.edu

Surendrabikram Thapa
Department of Computer Science, Virginia Tech Blacksburg, USA
sbt@vt.edu

Usman Naseem
College of Science and Engineering, James Cook University, Australia
usman.naseem@jcu.edu.au

Abstract

Scientific research relies heavily on the exchange of knowledge through citations in academic literature. In the domain of astrophysics, the precise classification of citation functions and the extraction of contextual information are critical for understanding the vast universe of research papers. This paper presents the system description for the WIESP 2023 FOCAL shared task. We introduce an automated approach that leverages state-of-the-art language models, including ALBERT, RoBERTa, BERT, and DistillBERT, to classify citation functions and extract context within astrophysical paragraphs. Our system combines paraphrasing and question-answering techniques to achieve accurate results. Through comprehensive experiments, we demonstrate the robustness of our approach, with ALBERT consistently delivering strong performance.

1 Introduction

Scientific research is a dynamic process fueled by the exchange of knowledge and ideas among researchers (Goodman and Royall, 1988; Ghosal et al., 2022; Tsunokake and Matsubara, 2022). In the context of scientific research, citations also serve as evidence and reference to past studies (Garfield et al., 1964). In the realm of astrophysics, the citation of existing literature plays a pivotal role in advancing our understanding of the universe Researchers rely on citations to establish the foundation of their work, compare results, and build upon previous discoveries. However, not all citations serve the same purpose. Some citations provide essential background knowledge, while others are used for comparison, validation, or to support specific claims within a research paper (Lauscher et al., 2022).

The citation graph is a foundational concept in scientific research, including astrophysics, where it plays a pivotal role in knowledge dissemination and discovery (Jurgens et al., 2018; Guo and Dai, 2022). This intricate network of references connects research papers, providing a basis for understanding, validation, and navigation within the vast and dynamic field of astrophysics literature. Citations serve as the foundation of knowledge, allowing researchers to establish context, validate findings, and trace the intellectual lineage of ideas (Cohan et al., 2019a). They also facilitate collaboration, highlight emerging trends, and aid in the navigation of extensive literature. Understanding the functions of citations is crucial in harnessing the full potential of the citation graph, and the FOCAL challenge at IJCNLP-AACL 2023 (Grezes et al., 2023) seeks to automate this classification, contributing to the advancement of astrophysical research and knowledge dissemination. Furthermore, as part of this challenge, we aim to identify not only the functions of citations but also the associated span of text in the paragraph that justifies these functions, enhancing the depth of understanding within astrophysical literature.

Moreover, recent advancements in language models (LMs) have provided exciting opportunities to tackle this challenge more effectively. These models, which are at the forefront of natural language processing (NLP), stand as powerful tools at the intersection of artificial intelligence and linguistics (Min et al., 2023; Thapa and Adhikari, 2023). Their growing capabilities, marked by their ability to understand and generate human-like text, present an opportunity to automate the classification of citation functions and the extraction of associated contextual information within the scientific literature.

In this paper, we introduce a comprehensive approach that leverages recent advancements in language models. Our methodology harnesses the power of paraphrasing and question-answering techniques to classify citation functions and extract relevant contextual spans within astrophysical paragraphs. We emphasize the adaptability and ver-

Proceedings of the 2nd Workshop on Information Extraction from Scientific Publications, pages 136–142
Nov 1, 2023.

satility of this approach, showcasing its potential applicability to various state-of-the-art language models. Through our efforts, we aim to contribute to the automation of citation function classification, ultimately advancing the accessibility and utility of astrophysics literature for researchers.

2 Task Description

The FOCAL (Function Of Citation in Astrophysics Literature) challenge (Grezes et al., 2023) presents a unique opportunity to delve into the intricate interplay between scientific literature and automated natural language processing.

2.1 Objective

Given a paragraph of text from the astrophysics literature, the challenge aims to develop machine learning models that can accurately determine why a citation is made in a given paragraph of astrophysics literature and identify the precise span of text within that paragraph that justifies the citation's function.

2.2 Dataset

The dataset provided for the FOCAL shared task consists of full-text fragments extracted from the NASA Astrophysics Data System (ADS) and has been meticulously annotated by domain experts to include essential information for the task.

Each entry in the dataset[1] for FOCAL 2023 adheres to the JSON Lines format, comprising a JSON dictionary with the following key elements:

- "Identifier": A unique string serving as an identifier for the entry, ensuring traceability and organization.
- "Paragraph": A text string extracted from astrophysics papers, which forms the basis for the citation function classification task.
- "Citation Text": A list of strings representing the citation(s) within the paragraph. While in most cases, this is a single string, there are instances where the citation text may be divided into multiple strings.

In the training dataset, the following additional information is provided:

- "Citation Start End": A list of integer pairs indicating the starting and ending positions of the citation(s) within the "Paragraph" text. In cases where the citation text is divided, multiple pairs are provided in corresponding order.
- "Functions Text": A list of strings highlighting portions of the paragraph that elucidate the function(s) of the citation(s). These strings serve as contextual evidence for understanding why the citation(s) were made.
- "Functions Label": A list of strings containing labels for each text element in "Functions Text." These labels correspond to the classification of the citation(s)' function(s) within the paragraph.
- "Functions Start End": A list of integer pairs indicating the starting and ending positions of the elements in "Functions Text" within the "Paragraph" text. Similar to the "Citation Start End" information, multiple pairs may exist when the "Functions Text" is divided.

> In some cases, when the pulse broadening time is a significant fraction of the pulse period (30 per cent or more) one can see a relatively sharp pulse, but at the same time the extended scattering tail may obscure the real baseline level, which leads to an underestimation of the pulsar flux. For pulsars with DMs in 200–300 pc cm-3 range this usually happens between 300 and 600 MHz (Lewandowski et al. 2013, 2015a). This leads to a somewhat pseudo-correlation between high DM and GPS pulsars (Kijak et al. 2007, 2011b) where serious underestimation of the flux at lower frequencies for high DM pulsars may give rise to an inverted spectra. The interferometric imaging technique provide a more robust measurement of the pulsar flux owing to the baseline lying at zero level thereby reducing errors made during the baseline subtraction.

As shown in the above paragraph, for the citation "Kijak et al. 2007" with start position = 495 and end position = 511, the expected model output is as follows:

- Function Labels: [Uses, Uses]
- Functions Start End: [(418, 492), (521, 640)]

[1] https://huggingface.co/datasets/adsabs/FOCAL

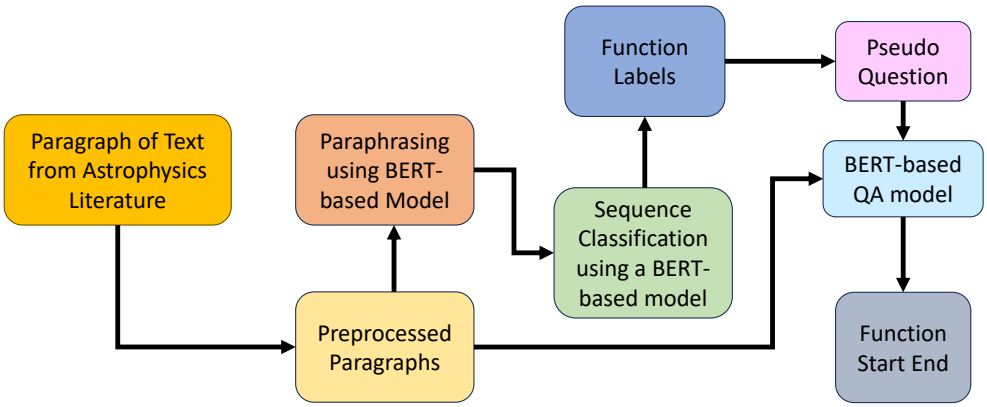

Figure 1: Our proposed approach for predicting citation function and associated span of text. We conduct tests with BERT, RoBERTa, DistillBERT, and ALBERT. A single language model (LM) is used for paraphrasing, sequence classification, and question-answering throughout the pipeline, resulting in four different configurations for the four models.

This output corresponds to the following textual evidence for the citation function:

Function Text:

- "This leads to a somewhat pseudo-correlation between high DM and GPS pulsars"
- "where serious underestimation of the flux at lower frequencies for high DM pulsars may give rise to an inverted spectra."

3 System Description

Our model leverages paraphrasing of the paragraphs and question answering for this task. Figure 1 shows the high-level overview of our model. We describe the system below:

3.1 Preprocessing of Paragraphs

We preprocess the text to input to our model. In the example paragraph shown above, we break them down into further parts based on the number of citations. For each citation, we take one fragment out of the paragraph. For each citation, we take the sentence in which the citation is up to the position where next citation starts. For Lewandowski et al. 2013, 2015a as shown above, we use the text as "*For pulsars with DMs in 200–300 pc cm-3 range this usually happens between 300 and 600 MHz (Lewandowski et al. 2013, 2015a). This leads to a somewhat pseudo-correlation between high DM and GPS pulsars*". Similarly, if the citation is the last one in the paragraph, we take the sentence in which a citation is in till the end of the paragraph. For Kijak et al. 2007, 2011b as shown above, we use the text as "*This leads to a somewhat pseudo-correlation between high DM and GPS pulsars (Kijak et al. 2007, 2011b) where serious underestimation of the flux at lower frequencies for high DM pulsars may give rise to an inverted spectra. The interferometric imaging technique provide a more robust measurement of the pulsar flux owing to the baseline lying at zero level thereby reducing errors made during the baseline subtraction.*" The preprocessed paragraphs are then fed into the paraphrasing model.

3.2 Language Models

Specifically, we use four BERT-based language models for paraphrasing, sequence classification, and QA model which are briefly described as follows.

BERT has achieved remarkable success in language understanding tasks by training on a massive amount of text data in a bidirectional manner, allowing it to understand the context and nuances of words and phrases (Devlin et al., 2019). This contextual understanding enables BERT to excel in a wide range of natural language understanding tasks, including text classification, question answering, and language translation (Papadopoulos et al., 2022; Zhou and Srikumar, 2022; Veeramani et al., 2023a,b,d,f). BERT's pre-trained embeddings have become a foundational resource in the world of natural language processing, serving as a starting point for various downstream tasks and research advancements (Adhikari et al., 2023).

RoBERTa is an acronym for "A Robustly Optimized BERT Pretraining Approach" (Liu et al.,

2019). It is a variant of the Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) model. RoBERTa builds upon the success of BERT by refining its pretraining methodology. It incorporates extensive training data, larger batch sizes, and longer training times, resulting in significantly improved performance on various natural language understanding tasks. RoBERTa is known for its robustness and exceptional performance on a wide range of text classification and language understanding tasks.

ALBERT is a model designed to reduce the computational and memory requirements of BERT while maintaining or even improving its performance (Lan et al., 2019). ALBERT achieves this by introducing parameter-sharing techniques, effectively reducing the model's size and training time. Despite its lighter architecture, ALBERT demonstrates remarkable efficiency and competitive performance across various natural language processing tasks (Kanagasabai et al., 2023). Its ability to handle large-scale text data with fewer computational resources makes it an appealing choice for resource-efficient applications.

DistillBERT is a distilled version of the original BERT model, emphasizing model compression and efficiency (Sanh et al., 2019). DistillBERT retains much of the performance of the larger BERT model while significantly reducing its size and computational requirements. This model distillation process involves training a smaller model (the "student") to mimic the behavior of a larger, more complex model (the "teacher"). DistillBERT is characterized by its compact size, making it suitable for deployment in resource-constrained environments without compromising accuracy.

3.3 Paraphrasing using BERT-based Model

In our approach, we leverage BERT-based models mentioned in section 3.2. BERT's contextual embeddings enable us to rephrase citation-related text effectively. We use paraphrasing in our pipeline in order to limit the input context to a length of 512.

3.4 Sequence Classification

Sequence classification serves as a fundamental component of our methodology (Cohan et al., 2019b; Veeramani et al., 2023c,e). We employ advanced language models mentioned in section 3.2 to classify the functions of citations within astrophysical paragraphs. This involves mapping citation-related segments to predefined categories, enabling us to clarify why each citation is made within the context of the research paper. The output is a multi-label output since a citation might be used for multiple purposes. The sequence classification component effectively outputs the "Functions Label".

3.5 Pseudo Question Generation

For each of the corresponding preprocessed text, we use their "Function Label" to form a pseudo question. This pseudo question serves as an input to the QA model. We form questions as "*What is the paragraph segment that corresponds to the function <FUNCTION LABEL>?*" For example, if we are looking for what part is background, our question is formed as "*What is the paragraph segment that corresponds to the function background?*"

3.6 BERT-based QA model

In our approach, we employ a BERT-based Question Answering (QA) model to further enhance the extraction of citation functions and their associated context. The QA model plays a pivotal role in our pipeline. The preprocessed text, as described in section 3.1, serves as one of the two inputs to our BERT-based QA model. This text contains the segmented paragraphs with citation-related information.

In our formulation, we formulate a pseudo question for each segment of the preprocessed text as the second input. This pseudo question is designed to encapsulate the essence of the citation function within the segment. It prompts the model to identify and extract the relevant information.

The output of our BERT-based QA model is a pair of integer values denoting the starting and ending positions of the citation function within the segment of text. These values pinpoint the exact location of the text that explains why the citation was made. We make the necessary adjustments for the offsets. By utilizing this QA model, we refine the precision and accuracy of our approach, providing explicit boundaries for the citation functions within the context of astrophysical paragraphs.

4 Results

The results presented in Table 1 demonstrate the performance of our approach utilizing various language models on the validation dataset for the FOCAL challenge. We evaluated our models

using three key metrics: 'seqeval_full', 'seqeval_generic', and 'labels_only'. Table

Model	seqeval_full	seqeval_generic	labels_only
BERT	0.2222	0.4393	0.4100
DistillBERT	0.2215	0.4369	**0.4985**
RoBERTa	0.2369	0.4356	0.4166
ALBERT	**0.2380**	**0.4396**	0.4261

Table 1: Performance of our approach with different language models on validation dataset

4.1 Validation Results

In terms of the 'seqeval_full' metric, which assesses the overall ability to correctly classify the functions of citations while ensuring accurate function labels, ALBERT achieved the highest score of 0.2380, closely followed by RoBERTa with a score of 0.2369. BERT and DistillBERT also performed reasonably well but exhibited slightly lower scores.

The 'seqeval_generic' metric, which evaluates the model's proficiency in identifying the portions of the paragraph that explain the functions of citations, showed a similar trend. ALBERT outperformed the other models with a score of 0.4396, followed closely by BERT, DistillBERT, and RoBERTa.

In terms of 'labels_only', which focuses solely on the accuracy of predicted function labels, DistillBERT led the pack with an F1-score of 0.4261, followed by ALBERT, RoBERTa, and BERT.

4.2 Test Results

In Table 2, we present the F1-score results on the test dataset using our approach with three different language models: BERT, RoBERTa, and ALBERT. The F1-scores are reported for three different evaluation metrics: seqeval_full, seqeval_generic, and labels_only.

seqeval_full Metrics: These metrics evaluate the overall ability to correctly classify the functions of citations while considering function labels.

- Micro F1-score: BERT achieved a micro F1-score of 0.27, RoBERTa scored 0.27, and ALBERT outperformed both with a micro F1-score of 0.30. Among the three, ALBERT shows the highest performance in this aspect.
- Macro F1-score: BERT scored the highest macro F1-score of 0.13, followed by RoBERTa (0.12) and ALBERT (0.12). BERT exhibits the highest average F1 score across different classes.
- Weighted F1-score: ALBERT achieves the highest weighted F1-score of 0.28, followed by BERT (0.28) and RoBERTa (0.28).

seqeval_generic Metrics: These metrics assess the model's proficiency in identifying portions of the paragraph that explain citation functions, regardless of the correctness of predicted function labels.

- Micro F1-score: ALBERT performs the best with a micro F1-score of 0.48, followed by RoBERTa (0.48) and BERT (0.47).
- Macro F1-score: ALBERT also achieves the highest macro F1-score of 0.48, while BERT and RoBERTa score similarly at 0.47 and 0.48 respectively.
- Weighted F1-score: ALBERT leads with a weighted F1-score of 0.48, followed by BERT (0.47) and RoBERTa (0.48).

labels_only Metrics: These metrics focus solely on the accuracy of predicted function labels, excluding the assessment of identified spans in the text.

- Micro F1-score: ALBERT outperforms the other models with a micro F1-score of 0.58, while BERT scores 0.48 and RoBERTa scores 0.48.
- Macro F1-score: BERT and RoBERTa have similar macro F1-scores of 0.24 and 0.22, respectively, while ALBERT scores lower at 0.21.
- Weighted F1-score: BERT achieves the highest weighted F1-score of 0.54, followed by ALBERT (0.53) and RoBERTa (0.53).

Overall, these results suggest that ALBERT consistently performs well across all three evaluation metrics, indicating its effectiveness in classifying citation functions and extracting contextual information within astrophysical literature. However, it's important to note that all models demonstrated reasonable performance, underscoring the viability of our approach across different language models.

5 Conclusions

In this paper, we have presented a comprehensive approach for automated citation function classification and context extraction in the domain of

Model	seqeval_full			seqeval_generic			labels_only		
	micro	macro	weighted	micro	macro	weighted	micro	macro	weighted
BERT	0.27	**0.13**	0.28	0.47	0.47	0.47	0.48	**0.24**	**0.54**
RoBERTa	0.27	0.12	0.28	0.48	0.48	0.48	0.48	0.22	0.53
ALBERT	**0.30**	0.12	**0.28**	**0.48**	**0.48**	**0.48**	**0.58**	0.21	0.53

Table 2: F1-score on the test dataset using our approach.

astrophysics literature. Leveraging advanced language models, including BERT, RoBERTa, ALBERT, and DistillBERT, our system showcases a robust pipeline that combines paraphrasing and question-answering techniques to achieve accurate and insightful results. Our experiments demonstrate the robustness of our approach, with ALBERT consistently performing well in classifying citation functions and extracting contextual information. However, all models exhibit reasonable performance, showcasing the adaptability of our system. In the future, we aim to refine our approach further, potentially incorporating more advanced models and techniques to enhance citation function classification and context extraction for deeper insights in astrophysical research.

Limitations

The limitations of this work include the potential challenges associated with accurately classifying citation functions within the nuanced landscape of astrophysical literature. Despite the effectiveness of our approach, the inherent complexity and subjectivity of citation functions may result in instances of misclassification or incomplete understanding. Finally, while we strive for generalizability, the specificities of astrophysical language and citation practices may limit the applicability of our approach to other scientific domains.

Ethics Statement

Our study adheres to principles of academic integrity, transparency, and respect for intellectual property rights. We have meticulously cited and credited all sources and data used in our work, ensuring due recognition for prior research contributions.

References

Surabhi Adhikari, Surendrabikram Thapa, Usman Naseem, Hai Ya Lu, Gnana Bharathy, and Mukesh Prasad. 2023. Explainable hybrid word representations for sentiment analysis of financial news. *Neural Networks*, 164:115–123.

Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady. 2019a. Structural scaffolds for citation intent classification in scientific publications. In *Proceedings of NAACL-HLT*, pages 3586–3596.

Arman Cohan, Iz Beltagy, Daniel King, Bhavana Dalvi, and Dan Weld. 2019b. Pretrained language models for sequential sentence classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3693–3699, Hong Kong, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Eugene Garfield, Irving H Sher, Richard J Torpie, et al. 1964. The use of citation data in writing the history of science.

Tirthankar Ghosal, Sergi Blanco-Cuaresma, Alberto Accomazzi, Robert M. Patton, Felix Grezes, and Thomas Allen, editors. 2022. *Proceedings of the first Workshop on Information Extraction from Scientific Publications*. Association for Computational Linguistics, Online.

Steven N Goodman and Richard Royall. 1988. Evidence and scientific research. *American Journal of Public Health*, 78(12):1568–1574.

Felix Grezes, Thomas Allen, Tirthankar Ghosal, and Sergi Blanco-Cuaresma. 2023. Function of citation in astrophysics literature (focal): Findings of the shared task. In *Proceedings of the 2nd Workshop on Information Extraction from Scientific Publications*, Online. Association for Computational Linguistics.

Lin Guo and Qun Dai. 2022. Graph clustering via variational graph embedding. *Pattern Recognition*, 122:108334.

David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. 2018. Measuring the evolution of a scientific field through citation frames. *Transactions of the Association for Computational Linguistics*, 6:391–406.

Rajaraman Kanagasabai, Saravanan Rajamanickam, Hariram Veeramani, Adam Westerski, and Kim Jung Jae. 2023. Semantists at ImageArg-2023: Exploring Cross-modal Contrastive and Ensemble Models for Multimodal Stance and Persuasiveness Classification. In *Proceedings of the 10th Workshop on Argument Mining*, Online and in Singapore. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Anne Lauscher, Brandon Ko, Bailey Kuehl, Sophie Johnson, Arman Cohan, David Jurgens, and Kyle Lo. 2022. MultiCite: Modeling realistic citations requires moving beyond the single-sentence single-label setting. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1875–1889, Seattle, United States. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2):1–40.

Christos Papadopoulos, Yannis Panagakis, Manolis Koubarakis, and Mihalis Nicolaou. 2022. Efficient learning of multiple NLP tasks via collective weight factorization on BERT. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 882–890, Seattle, United States. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Surendrabikram Thapa and Surabhi Adhikari. 2023. Chatgpt, bard, and large language models for biomedical research: Opportunities and pitfalls. *Annals of Biomedical Engineering*, pages 1–5.

Masaya Tsunokake and Shigeki Matsubara. 2022. Classification of URL citations in scholarly papers for promoting utilization of research artifacts. In *Proceedings of the first Workshop on Information Extraction from Scientific Publications*, pages 8–19, Online. Association for Computational Linguistics.

Hariram Veeramani, Surendrabikram Thapa, and Usman Naseem. 2023a. DialectNLU at NADI 2023 Shared Task: Transformer Based MultiTask Approach Jointly Integrating Dialect and Machine Translation Tasks. In *Proceedings of The First Arabic Natural Language Processing Conference (ArabicNLP 2023)*, Singapore. Association for Computational Linguistics.

Hariram Veeramani, Surendrabikram Thapa, and Usman Naseem. 2023b. Enhancing ESG Impact Type Identification through Early Fusion and Multilingual Models. In *Proceedings of the Sixth Workshop on Financial Technology and Natural Language Processing (FinNLP)*. Association for Computational Linguistics.

Hariram Veeramani, Surendrabikram Thapa, and Usman Naseem. 2023c. KnowTellConvince at ArAIEval 2023: Disinformation and Persuasion Detection using Similar and Contrastive Representation Alignment. In *Proceedings of The First Arabic Natural Language Processing Conference (ArabicNLP 2023)*, Singapore. Association for Computational Linguistics.

Hariram Veeramani, Surendrabikram Thapa, and Usman Naseem. 2023d. LowResContextQA at Qur'an QA 2023 Shared Task: Temporal and Sequential Representation Augmented Question Answering Span Detection. In *Proceedings of The First Arabic Natural Language Processing Conference (ArabicNLP 2023)*, Singapore. Association for Computational Linguistics.

Hariram Veeramani, Surendrabikram Thapa, and Usman Naseem. 2023e. Lowresourcenlu at blp: Enhancing sentiment classification and violence incitement detection through aggregated language models. In *Proceedings of the 1st International Workshop on Bangla Language Processing (BLP-2023)*, Singapore. Association for Computational Linguistics.

Hariram Veeramani, Surendrabikram Thapa, and Usman Naseem. 2023f. Temporally Dynamic Session-Keyword Aware Sequential Recommendation system. In *2023 International Conference on Data Mining Workshops (ICDMW)*.

Yichu Zhou and Vivek Srikumar. 2022. A closer look at how fine-tuning changes BERT. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1046–1061, Dublin, Ireland. Association for Computational Linguistics.