Enhancing ESG Impact Type Identification through Early Fusion and Multilingual Models

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

In the evolving landscape of Environmental, Social, and Corporate Governance (ESG) impact assessment, the ML-ESG-2 shared task proposes identifying ESG impact types. To address this challenge, we present a comprehensive system leveraging ensemble learning techniques, capitalizing on early and late fusion approaches. Our approach employs four distinct models: mBERT, FlauBERT-base, ALBERT-base-v2, and a Multi-Layer Perceptron (MLP) incorporating Latent Semantic Analysis (LSA) and Term Frequency-Inverse Document Frequency (TF-IDF) features. Through extensive experimentation, we find that our early fusion ensemble approach, featuring the integration of LSA, TF-IDF, mBERT, FlauBERT-base, and ALBERT-base-v2, delivers the best performance. Our system offers a comprehensive ESG impact type identification solution, contributing to the responsible and sustainable decision-making processes vital in today's financial and corporate governance landscape.

1 Introduction

In the rapidly evolving landscape of finance and corporate governance, Environmental, Social, and Corporate Governance (ESG) considerations have gained unprecedented prominence (Linhares Pontes et al., 2022). Investors, stakeholders, and businesses increasingly recognize the importance of ESG factors in decision-making processes, portfolio management, and corporate strategy (Li et al., 2021; Gillan et al., 2021). The "E" in ESG shines a spotlight on environmental sustainability, compelling companies to address pressing issues such as climate change and resource conservation. Recognizing that eco-friendly practices are ethical and financially savvy, businesses are embracing environmental responsibility as a key driver of long-term success (Alsayegh et al., 2020; Newell and Marzuki, 2022). Simultaneously, the "S" in ESG emphasizes social responsibility, pushing corpora-

tions to foster diversity, inclusivity, and community engagement (Buallay, 2019). Companies that prioritize these social factors not only enhance their reputation but also attract top talent and build resilience in a socially conscious world. Finally, the "G" underscores the importance of sound corporate governance, providing the foundation for trust and stability in financial markets. Investors now realize that transparent decision-making, ethical conduct, and effective risk management are essential elements for ensuring a company's long-term viability (Tian et al., 2022).

In this context, the ML-ESG-2 shared task is pivotal in facilitating a deeper understanding of how ESG information can be extracted and utilized from the vast sea of textual data, contributing to the broader endeavor of responsible and sustainable finance. This specific task of ESG impact type identification, where machine learning models classify news articles or sentences as "Opportunity", "Risk", or "Cannot Distinguish", carries profound significance. For investors, this knowledge translates into well-informed decisions with the potential to impact portfolios significantly (Amel-Zadeh and Serafeim, 2018). Positive ESG signals can guide investments toward companies poised for sustainable growth while identifying ESG risks enables proactive risk mitigation. In the corporate sphere, grasping the ESG implications of textual data like news articles and annual reports fosters strategic decision-making and risk management. Companies can leverage positive ESG news to enhance their sustainability efforts and attract responsible investors while swiftly addressing negative news to minimize reputational damage and regulatory challenges (Goel et al., 2022). Furthermore, automated text analysis tools expedite these processes, enabling scalable and efficient ESG information extraction.

As global issues like climate change, social inequality, and ethical governance become increas-

ingly prominent, harnessing ESG information from textual sources is a means to hold corporations accountable. By scrutinizing relevant text like news articles or annual securities reports for their ESG impact, stakeholders can encourage responsible behavior, promote sustainable practices, and ensure that businesses prioritize societal and environmental concerns (Kannan and Seki, 2023). This task not only empowers decision-makers with the ability to navigate the complexities of ESG information but also contributes to a broader cultural shift towards responsible investment and corporate citizenship. In essence, ESG information extraction from text represents a vital step towards a more sustainable and equitable future, where financial decisions align with environmental and social responsibility goals.

In Natural Language Processing (NLP), text classification is a critical enabler for tasks like ESG impact type identification, bridging the gap between the vast expanse of textual data and actionable insights. NLP techniques have been widely used in the financial domain (Jaggi et al., 2021; Adhikari et al., 2023). Leveraging NLP techniques, machine learning models can autonomously categorize news articles or sentences into predefined ESG impact types, such as "Opportunity", "Risk", or "Cannot Distinguish". These models harness the power of feature extraction, supervised learning, feature engineering, and even advanced transfer learning from pre-trained language models to accurately assess the implications of textual content on a company's ESG performance. The synergy between NLP and ESG, impact type identification, is instrumental in empowering stakeholders, including investors and corporations, to efficiently navigate the complexities of ESG information, make informed decisions, and contribute to a more sustainable and responsible financial and corporate governance landscape (Kang and El Maarouf, 2022).

This paper presents the system description of our submission to ML-ESG-2, a shared task hosted by the FinNLP workshop in conjunction with IJCNLP-AACL 2023. In this task, we delve into the critical intersection of Natural Language Processing (NLP) and ESG impact type identification.

2 Task Description

ML-ESG-2 introduces a new task called ESG impact type identification.

2.1 Objective

Given textual data, the objective is to classify it into one of three categories from the ESG aspect: "Opportunity", "Risk", or "Cannot Distinguish" (or "Positive", "Negative", or "N/A" in the Japanese dataset).

2.2 Dataset

The ML-ESG-2 task utilizes a carefully curated dataset by Kannan and Seki (2023) to support the ESG impact type identification objective. This dataset was curated using Japanese annual securities reports as a primary source of information to extract insights into a company's ESG (Environmental, Social, and Corporate Governance) initiatives. In the context of ESG impact type identification, each news article or sentence is categorized into one of three distinct labels. The label "Opportunity" is assigned to articles or sentences that suggest a positive impact or opportunity for the company from an ESG perspective. Conversely, the label "Risk" is assigned to articles or sentences that convey a negative impact or potential risk to the company concerning ESG factors. However, in cases where the available content does not provide sufficient clarity to definitively classify the article as either an opportunity or a risk, the label "Cannot Distinguish" is used. It is worth noting that in the Japanese dataset, these labels are represented as "Positive", "Negative", and "N/A" to align with the language and terminology commonly used in Japanese ESG contexts.

3 System Description

In this section, we provide a detailed overview of our system architecture and methodology for the ML-ESG-2 shared task on ESG impact type identification. Our approach is founded on the principles of ensemble learning, utilizing both early fusion and late fusion techniques to harness the strengths of multiple language models and feature representations. In order to approach this task, we use four different models.

3.1 Models Used

In our system, we leverage the predictive capabilities of four different models, each contributing unique strengths to the ESG impact type identification process:
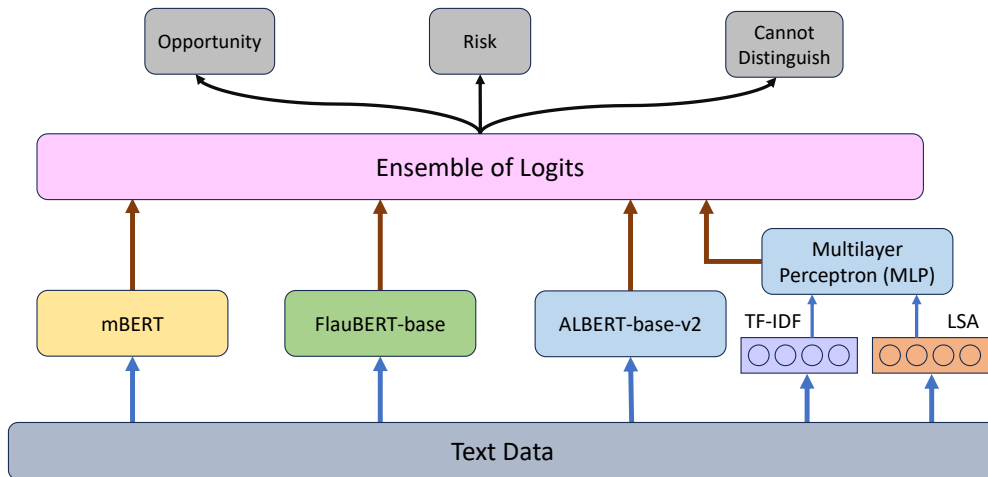
Figure 1: Late fusion technique that uses logits from all models to make the final decision.

3.1.1 mBERT (Multilingual BERT)

To ensure robust multilingual text analysis, we fine-tune mBERT (Devlin et al., 2019; Wu and Dredze, 2020) on the task-specific dataset. This model adapts to contextual information and linguistic nuances in the news articles, making it a valuable component for handling multilingual content (Veeramani et al., 2023c,b,d) effectively.

3.1.2 FlauBERT-base (French Language Model)

Given the presence of French textual content in the task, we incorporate FlauBERT-base (Le et al., 2020), a specialized French language model. This addition ensures accurate analysis and classification of French content, enhancing the overall system's robustness.

3.1.3 ALBERT-base-v2

We integrate ALBERT-base-v2 (Lan et al., 2019), known for its efficient parameterization and performance, to diversify our model components. This model further enhances our system's capability to capture nuanced ESG-related nuances and impact types.

3.1.4 MLP (Multi-Layer Perceptron)

Complementing the language models, we employ a Multi-Layer Perceptron (MLP) (Murtagh, 1991) that incorporates feature representations derived from Latent Semantic Analysis (LSA) (Dumais, 2004) and Term Frequency-Inverse Document Frequency (TF-IDF) (Bafna et al., 2016; Adhikari et al., 2021). These features capture semantic infor-

mation and term importance within the documents. The MLP facilitates the modeling of intricate feature interactions (Veeramani et al., 2023a,e), enhancing the depth of our system's analysis.

3.2 Ensemble Techniques

Our ensemble strategy combines both early fusion and late fusion techniques, capitalizing on the collective intelligence of individual models and feature representations.

3.2.1 Late Fusion Ensemble

In the late fusion approach, we aggregate predictions from individual models and feature representations at the logits level (Kanagasabai et al., 2023). This technique effectively combines the diverse outputs generated by mBERT, FlauBERT-base, ALBERT-base-v2, and the MLP with LSA and TF-IDF features. Late fusion ensures comprehensive integration of the predictive capabilities of these components, resulting in a more robust and accurate prediction for ESG impact type identification. Figure 1 shows the schematic overview of the late fusion ensembling technique we employ.

3.2.2 Early Fusion Ensemble

In the early fusion approach, we integrate representations from mBERT, FlauBERT-base, ALBERT-base-v2, LSA, and TF-IDF features at an earlier stage, prior to inputting them into the MLP for prediction. Early fusion allows for seamless information integration from multiple sources, empowering the MLP to capture intricate relationships among these elements. This approach ensures the
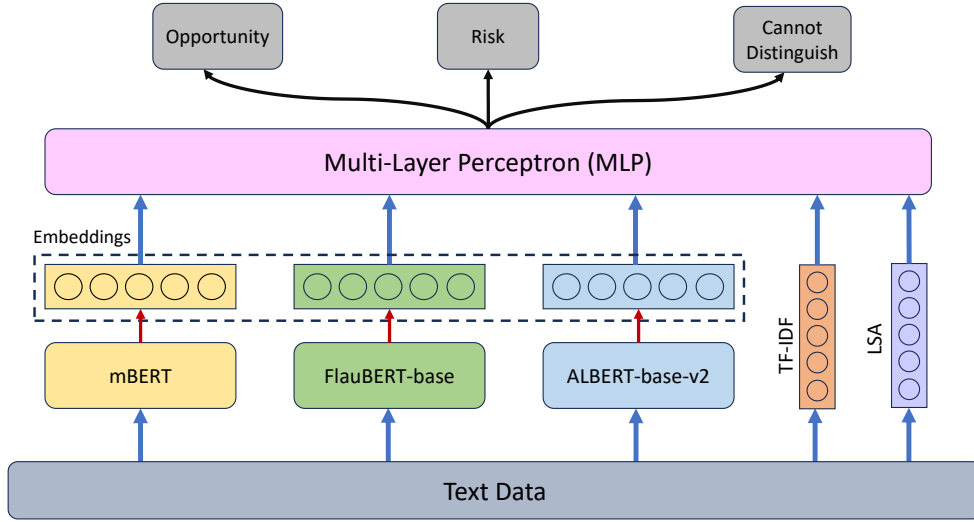
Figure 2: Early fusion ensemble takes the different representations and uses MLP for the final classification.

Embeddings	Language	Micro-F1	Macro-F1	Weighted-F1
FlauBERT + mBERT + ALBERT + TF-IDF	English	0.9587	0.9128	0.9587
	French	0.545	0.520	0.5111
	Japanese	0.5323	0.2933	0.4905
	Chinese	0.403	0.1667	0.3782
FlauBERT + mBERT + ALBERT + TF-IDF + LSA	English	0.9633	0.918	0.9639
	French	0.5501	0.5292	0.5155
	Japanese	0.5378	0.3043	0.4943
	Chinese	0.4102	0.1728	0.3881

Table 1: Ablation study for different combinations of embeddings with MLP layers as the classification layer (early fusion).

extraction of richer semantic and contextual features, enhancing the system's ability to make nuanced predictions for identifying ESG impact types in news articles. Figure 2 shows the schematic overview of the early fusion ensembling technique we employ.

By adopting both early and late fusion ensemble techniques, our system maximizes accuracy and robustness, offering a powerful solution for the task of ESG impact type identification in textual data.

4 Results

In the ML-ESG-2 shared task for ESG impact type identification, we evaluated the performance of our system across multiple languages using different ensemble configurations and fusion techniques. The micro F1, macro F1, and weighted F1 scores offer a comprehensive assessment of our system's

effectiveness in capturing nuances within different linguistic contexts.

Our system secured notable rankings in the shared task, claiming the fourth position for English, the third position for both Japanese and Chinese, and the fifth position for the French language. We employed various ensemble configurations, including late fusion models combining FLAUBERT, mBERT, and ALBERT, incorporating TF-IDF and Latent Semantic Analysis (LSA) features. The results demonstrate significant improvements when transitioning to early fusion techniques, particularly when combining TF-IDF, LSA, and language model representations, as shown in Table 1. The late fusion techniques, adding more information with multiple models, seemed to help, as shown in Table 2.

In the English language, our best early fusion ensemble achieved a micro F1 score of 0.9633, a

Models	Language	Micro-F1	Macro-F1	Weighted-F1
FlauBERT + mBERT + ALBERT	English	0.9403	0.8899	0.9357
	French	0.525	0.501	0.4933
	Japanese	0.5155	0.2755	0.465
	Chinese	0.378	0.1346	0.3525
FlauBERT + mBERT + ALBERT + MLP (TF-IDF)	English	0.945	0.8944	0.9403
	French	0.530	0.505	0.5022
	Japanese	0.520	0.280	0.475
	Chinese	0.384	0.141	0.3589
FlauBERT + mBERT + ALBERT + MLP (TF-IDF + LSA)	English	0.9495	0.8991	0.9449
	French	0.535	0.512	0.5066
	Japanese	0.5244	0.2899	0.480
	Chinese	0.391	0.1474	0.3717

Table 2: Ablation study for different combinations of models (late fusion).

Team	Micro-F1	Macro-F1	Weighted-F1
AnakItik	0.9817	0.9548	0.9810
BrothFink	0.9771	0.9445	0.9765
NeverCareU	0.9633	0.9227	0.9648
FinNLU	0.9633	0.9180	0.9639
231	0.9633	0.9127	0.9627
SPEvFT	0.9587	0.9118	0.9602
LIPI	0.9312	0.8335	0.9294
HHU	0.9174	0.8098	0.9174

Table 3: Leaderboard for ESG identification in English language

macro F1 score of 0.918, and a weighted F1 score of 0.9639, showcasing its proficiency in ESG impact type identification. Similarly, for other languages, the early fusion ensemble, which integrates TF-IDF and LSA features with language models, achieved the highest performance among all early fusion and late fusion techniques.

While our ensemble system demonstrates commendable performance in English (as shown in Table 3), it is evident that its effectiveness diminishes when applied to other languages such as French, Japanese, and Chinese. This discrepancy underscores the importance of adapting our approach to address language-specific nuances and complexities. To remedy this, future research efforts may focus on language-specific model fine-tuning, dataset augmentation with more diverse linguistic content, and the development of specialized language models tailored to the unique challenges posed by each language. Additionally, incorporating more extensive language-specific features and linguistic resources into our ensemble configurations could further enhance the system's adaptability and robustness across diverse linguistic contexts. These measures hold the potential to improve our system's performance and make it a valuable tool for multilingual ESG impact type identification.

5 Conclusion

In the rapidly evolving finance and corporate governance landscape, the ML-ESG-2 shared task has presented a unique opportunity to explore ESG impact type identification. Our comprehensive system leverages the collective intelligence of multiple language models and feature representations through advanced ensemble learning techniques, resulting in high performance. While our system excels in English, it faces hurdles in accurately identifying ESG impact types in Japanese, Chinese, and French languages. These challenges underscore the importance of further research and fine-tuning to adapt our approach to diverse linguistic contexts.

Limitations

While our ensemble learning approach has demonstrated strong performance, we acknowledge certain limitations in our study. The system's performance variation across languages, with lower accuracy for Japanese, Chinese, and French texts, highlights the need for further language-specific model development and fine-tuning. Additionally, the reliance on single-source textual data alone may not capture all relevant context and nuances, which may affect the system's ability to identify ESG impact types accurately in certain cases. Furthermore,

the generalization of our approach to other ESG-related tasks and domains may require additional adaptation and validation. We also recognize the evolving nature of language models and the need for continuous updates and refinements to maintain their effectiveness.

Ethics Statement

Throughout our research, we recognize that ethical considerations, particularly those related to bias mitigation and fairness in model predictions, represent ongoing and critical areas for improvement in the development of responsible AI systems. It is important to note that, in this study, we have not undertaken specific measures to actively mitigate bias within our models or predictions.

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.

Surabhi Adhikari, Surendrabikram Thapa, Priyanka Singh, Huan Huo, Gnana Bharathy, and Mukesh Prasad. 2021. A comparative study of machine learning and nlp techniques for uses of stop words by patients in diagnosis of alzheimer's disease. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Maha Faisal Alsayegh, Rashidah Abdul Rahman, and Saeid Homayoun. 2020. Corporate economic, environmental, and social sustainability performance transformation through esg disclosure. *Sustainability*, 12(9):3910.

Amir Amel-Zadeh and George Serafeim. 2018. Why and how investors use esg information: Evidence from a global survey. *Financial analysts journal*, 74(3):87–103.

Prafulla Bafna, Dhanya Pramod, and Anagha Vaidya. 2016. Document clustering: Tf-idf approach. In *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, pages 61–66. IEEE.

Amina Buallay. 2019. Is sustainability reporting (esg) associated with performance? evidence from the european banking sector. *Management of Environmental Quality: An International Journal*, 30(1):98–115.

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.

Susan T Dumais. 2004. Latent semantic analysis. *Annual Review of Information Science and Technology (ARIST)*, 38:189–230.

Stuart L Gillan, Andrew Koch, and Laura T Starks. 2021. Firms and social responsibility: A review of esg and csr research in corporate finance. *Journal of Corporate Finance*, 66:101889.

Tushar Goel, Vipul Chauhan, Suyash Sangwan, Ishan Verma, Tirthankar Dasgupta, and Lipika Dey. 2022. TCS WITM 2022@FinSim4-ESG: Augmenting BERT with linguistic and semantic features for ESG data classification. In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*, pages 235–242, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Mukul Jaggi, Priyanka Mandal, Shreya Narang, Usman Naseem, and Matloob Khushi. 2021. Text mining of stocktwits data for predicting stock prices. *Applied System Innovation*, 4(1):13.

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.

Juyeon Kang and Ismail El Maarouf. 2022. FinSim4-ESG shared task: Learning semantic similarities for the financial domain. extended edition to ESG insights. In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*, pages 211–217, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Naoki Kannan and Yohei Seki. 2023. Textual evidence extraction for ESG scores. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, pages 45–54, Macao. -.

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*.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. 2020. Flaubert: Unsupervised language model pre-training for french. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2479–2490.

Ting-Ting Li, Kai Wang, Toshiyuki Sueyoshi, and Derek D Wang. 2021. Esg: Research progress and future prospects. *Sustainability*, 13(21):11663.

Elvys Linhares Pontes, Mohamed Ben Jannet, Jose G. Moreno, and Antoine Doucet. 2022. Using contextual sentence analysis models to recognize ESG concepts. In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*, pages 218–223, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Fionn Murtagh. 1991. Multilayer perceptrons for classification and regression. *Neurocomputing*, 2(5-6):183–197.

Graeme Newell and Muhammad Jufri Marzuki. 2022. The increasing importance of environmental sustainability in global real estate investment markets. *Journal of Property Investment & Finance*, 40(4):411–429.

Ke Tian, Zepeng Zhang, and Hua Chen. 2022. Automatic Term and Sentence Classification Via Augmented Term and Pre-trained language model in ESG Taxonomy texts. In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*, pages 224–227, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Hariram Veeramani, Surendrabikram Thapa, and Usman Naseem. 2023a. Automated Citation Function Classification and Context Extraction in Astrophysics: Leveraging Paraphrasing and Question Answering. In *Proceedings of the second Workshop on Information Extraction from Scientific Publications*, Online. Association for Computational Linguistics.

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

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual bert? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130.