KnowTellConvince at ArAIEval Shared Task: Disinformation and Persuasion Detection in Arabic using Similar and Contrastive Representation Alignment

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 an era of widespread digital communication, the challenge of identifying and countering disinformation has become increasingly critical. However, compared to the solutions available in the English language, the resources and strategies for tackling this multifaceted problem in Arabic are relatively scarce. To address this issue, this paper presents our solutions to tasks in ArAIEval 2023. Task 1 focuses on detecting persuasion techniques, while Task 2 centers on disinformation detection within Arabic text. Leveraging a multi-head model architecture, fine-tuning techniques, sequential learning, and innovative activation functions, our contributions significantly enhance persuasion techniques and disinformation detection accuracy. Beyond improving performance, our work fills a critical research gap in content analysis for Arabic, empowering individuals, communities, and digital platforms to combat deceptive content effectively and preserve the credibility of information sources within the Arabic-speaking world.

1 Introduction

In today's information age, the rapid dissemination of digital content across various platforms has revolutionized the way information is produced, shared, and consumed. This unprecedented accessibility to information has brought numerous benefits, but it has also given rise to new challenges, particularly in the realms of misinformation, propaganda, and disinformation (Alam et al., 2022a). Identifying and addressing these issues is paramount for ensuring the integrity and credibility of information sources.

While the English language has garnered substantial attention in the realm of misinformation, persuasion, and disinformation detection, it is imperative that we recognize the equal, if not greater, significance of these endeavors in the Arabic language. Less research on these areas will leave the Arabic-speaking world vulnerable to the harmful effects of deceptive content. Arabic's linguistic and cultural nuances demand tailored approaches to combat these issues effectively (Sheikh Ali et al., 2023; Alyoubi et al., 2023; Fouad et al., 2022).

Disinformation, encompassing hate speech, offensive content, rumors, spam, and propaganda, presents formidable challenges in the Arabic-speaking world, shaped by linguistic diversity and cultural nuances (Nakov et al., 2022; Alam et al., 2022b). Hate speech and offensive content, intensified by cultural sensitivities, demand effective detection and mitigation to avert real-world repercussions (Albadi et al., 2018; Al-Hassan and Al-Dossari, 2022; Chowdhury et al., 2019). Rumors, highly contagious within tight-knit Arabic communities, necessitate vigilant monitoring to counteract panic and misinformation, exploiting cultural contexts for added complexity (Nakov et al., 2021; Harrag and Djahli, 2022). Spam, spanning fraudulent ads and misleading claims, pervades digital spaces in all languages, underlining the need to distinguish it from credible content for online source credibility (Kaddoura et al., 2023; Alkadri et al., 2022). Propaganda, a pivotal element of disinformation campaigns, influences public opinion and necessitates understanding and countering within the Arabic-speaking context to protect individuals and communities from manipulation by misleading narratives (Sharara et al., 2022; Feldman et al., 2021). Addressing these multifaceted challenges requires comprehensive research efforts and robust detection models that account for linguistic and cultural intricacies, preserving the credibility of information sources, online discourse, and public opinion in the diverse and dynamic Arabic-speaking linguistic landscape.

In order to address the problems mentioned above and to extend the previous related works (Habernal et al., 2017, 2018; Da San Martino et al., 2019; Barrón-Cedeno et al., 2019), in this paper, we

Proceedings of the The First Arabic Natural Language Processing Conference (ArabicNLP 2023), pages 519–524
December 7, 2023

present a multi-faceted approach that combines innovative model architectures, fine-tuning strategies, and sequential learning techniques to effectively address subtask 1A of Task 1 (persuasion or propaganda detection) and both subtasks of Task 2 (disinformation detection) in ArAIEval 2023 (Hasanain et al., 2023). Our incorporation of contrastive learning, renormalization, sentence embedding, cosine similarity checks, and GELU activation functions within the Arabic BERT framework demonstrates a comprehensive strategy for detecting disinformation subtleties, including hate speech, offensive content, rumors, spam, and propaganda. Our contributions not only enhance disinformation detection accuracy but also bridge the research gap in content analysis for the Arabic language.

2 Task Description

Task 1: This task mainly deals with persuasion techniques (propagandistic content) and has two subtasks. Our participation is focused on subtask 1A, which involves analyzing individual paragraphs of text from various genres to determine whether they contain persuasive content, with a binary classification of "Yes" or "No" as the output.

Task 2: This task centers on disinformation detection with two subtasks: subtask 2A for binary classification to identify disinformation in tweets and subtask 2B for multi-class classification, categorizing tweets into hate speech, offensive content, rumors, or spam categories.

3 Dataset

Task 1: The dataset comprises tweets and news paragraphs that have been annotated to identify the use of persuasion techniques. These annotations are provided in binary and multilabel settings, allowing for the classification of the presence or absence of persuasion techniques and, in the multilabel setting, the identification of multiple propaganda techniques within the same text. Since we only participate in subtask 1A, we only use binary annotation data. The development set contains approximately 78% of the data without propaganda and 22% of the data with propaganda. Similarly, the test set comprises roughly 34.2% of the data without propaganda and 65.8% of the data with propaganda.

Task 2: Similar to Task 1, this task also contains tweets annotated for binary and multiclass labels for subtask 2A and subtask 2B, respectively.

4 System Descriptions

Our system is an ensemble of four models, as shown in Figure 1. Below, we explain every component in detail.

Model A - Supervised Contrastive Learning with Arabic BERT: In Model A, we employ contrastive learning to enhance Arabic text representations. The motivation is to empower the model for binary classification tasks by improving its ability to distinguish between positive and negative examples in Arabic text (Alam et al., 2022b; Veeramani et al., 2023b,d,c). Contrastive learning encourages the model to capture semantic relationships effectively, benefiting applications like sentiment analysis. We fine-tune BERT Arabic Base (Safaya et al., 2020) with a contrastive loss function, pushing the model to generate embeddings emphasizing semantic similarity and dissimilarity. During training, it promotes similar representations for similar sentences and different representations for dissimilar sentences, enhancing the model's semantic understanding.

Model B - Sequential Learning with ArabicBERT: Model B adopts a sequential learning approach, fine-tuning ArabicBERT on task-specific data to adapt to various Arabic NLP tasks. The motivation is to enable the model to comprehend sequential relationships and context in textual data, which is crucial for tasks like text generation and named entity recognition. Rationally, sequential learning involves taking the pretrained BERT model and fine-tuning it on specific tasks, transferring knowledge from its general language understanding capabilities to task-specific nuances. We adjust learning rates and batch sizes for different tasks and employ task-specific loss functions during fine-tuning.

Model C - Fine-Tuned Arabic BERT with Renormalized XNLI Data and Sequential Learning: This process entails taking the pretrained Arabic BERT model and fine-tuning it on the renormalized XNLI dataset. During this phase, the labels in the dataset are adjusted to reflect the sentiment perspective, where neutral and entailment labels are unified into one class, and contradiction remains as a separate class. Subsequently, applying sequential learning further enhances the model's adaptation to the task-specific nuances (Gururangan et al.,

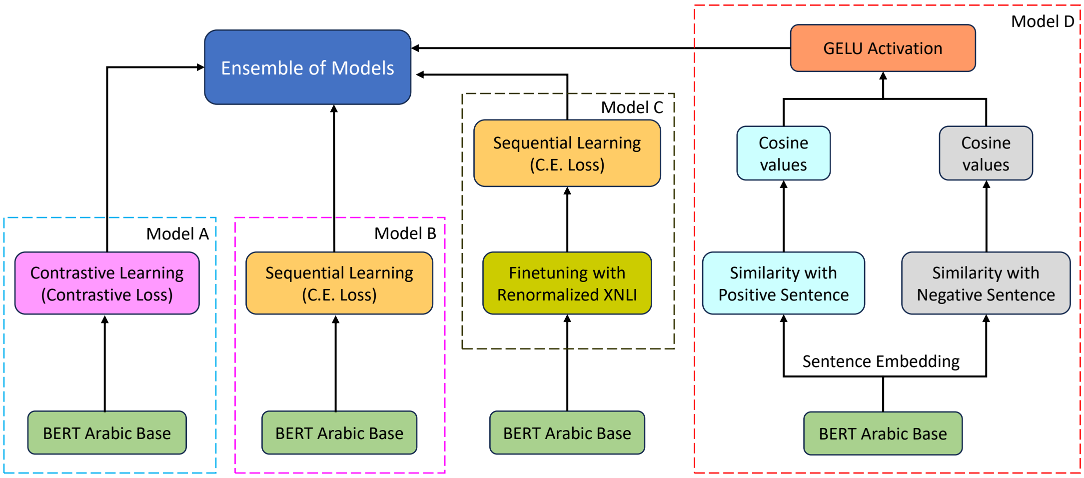

Figure 1: Overall framework of our proposed methodology.

2020; Da San Martino et al., 2019; Veeramani et al., 2023e,a,f). During sequential learning, the model adjusts its internal representations based on the fine-tuned XNLI data, further refining its understanding of sentiment-related features and patterns unique to Arabic text. This sequential fine-tuning ensures that the model aligns precisely with the propaganda/disinformation analysis and classification requirements.

Model D - Sentence Embeddings with GELU Activation: The process begins with calculating semantic similarity between sentences in Arabic text, serving as a foundational step for tasks demanding an understanding how closely related or similar two sentences are. The primary motivation behind this approach is to excel in applications that rely on measuring the semantic similarity between sentences in Arabic, such as persuasion detection, disinformation detection, and multiclass classification (Kanagasabai et al., 2023). The model extracts sentence embeddings from Arabic BERT representations to capture the essential features and semantics of each sentence. Subsequently, it calculates the cosine similarity between pairs of sentences, providing a quantitative measure of their semantic relatedness. To capture complex and non-linear relationships within the sentence embeddings, the model then applies the GELU (Gaussian Error Linear Unit) activation function. This step enhances the model's ability to discern intricate semantic nuances. Ultimately, the GELU-activated cosine similarity scores enable the model to assess the degree of semantic similarity between sentences, making Model D a valuable asset for tasks like persuasion and disinformation detection, which requires semantic understanding and similarity assessment in Arabic text processing.

5 Results and Discussion

This section discusses the results of our runs. Apart from the above mentioned, we also tested to ArSAS BERT[1]. We perform a detailed ablation of what factors contribute to the better performance of the system.

5.1 Task 1A (Persuasion Detection)

In the context of persuasion detection, the presented Table 1 reveals a comprehensive evaluation of various models designed to excel in this task. Arabic-BERT demonstrated commendable effectiveness with a micro-averaged F1-score of 72.23, emphasizing its proficiency in classifying instances. ArSAS BERT slightly outperformed Arabic-BERT with a micro-averaged F1-score of 73.4, highlighting its capabilities in persuasion detection. However, the combination of components B and D notably improved model performance, resulting in Model (B + D) achieving a micro-averaged F1-score of 74.44. Model (A + B) further enhanced performance to a micro-averaged F1-score of 74.75 by combining components A and B, showcasing the value of ensemble models. Nevertheless, the Model (A + D) also boasted a micro-averaged F1-score of 75.77, emphasizing the effectiveness of combining components A and D. The most comprehensive approach, Model (A + B + C + D),

[1] https://huggingface.co/Osaleh/sagemaker-bert-base-arabic-ArSAS

Models	$F1_{mic}$	Pre_{mac}	Rec_{mac}	$F1_{mac}$
Arabic-BERT	72.23	73.51	72.06	71.0
ArSAS BERT	73.4	73.17	74.8	73.2
Model (B + D)	74.44	75.09	74.58	74.67
Model (A + B)	74.75	75.48	74.75	75.05
Model (A + D)	75.77	76.9	75.26	76.05
Model (A+B+C+D)	76.14	78.11	76.14	76.82

Table 1: Results for task 1A (persuasion detection). The $F1_{mic}$ stands for micro-averaged F1-score. Similarly, Pre_{mac}, Rec_{mac}, and $F1_{mac}$ represents macro-averaged precision, recall and F1-score.

achieved the highest micro-averaged F1-score at 76.14, reaffirming the synergy of all four components in tackling persuasion detection effectively. These results underscore the significance of model combinations and component choices in optimizing performance for this task.

5.2 Task 2A (Disinformation Detection)

In disinformation detection, the provided Table 2 showcases a comprehensive evaluation of diverse models. Arabic-BERT demonstrated strong performance with a micro-averaged F1-score of 86.4, underscoring its effectiveness in identifying disinformation. ArSAS BERT improved upon this, achieving a micro-averaged F1-score of 87.26, signifying its proficiency in detecting false information. However, the strategic combination of components B and D notably enhanced model performance, resulting in Model (B + D) achieving a micro-averaged F1-score of 88.5. Model (A + B) excelled further with an impressive micro-averaged F1-score of 89.05, indicating its strength in disinformation detection. However, Model (A + D) emerged as a better performer, boasting a micro-averaged F1-score of 89.38 and demonstrating its exceptional capability in detecting disinformation. The most encompassing approach, Model (A + B + C + D), outshone the rest with the highest micro-averaged F1-score at 89.67, reaffirming the synergy of all four components in effectively combatting disinformation.

Models	$F1_{mic}$	Pre_{mac}	Rec_{mac}	$F1_{mac}$
Arabic-BERT	86.4	87	86.22	86.32
ArSAS BERT	87.26	88.5	87.15	87.2
Model (B + D)	88.5	89.02	88.46	88.9
Model (A + B)	89.05	89.88	89.06	89.35
Model (A + D)	89.38	90	89.38	89.61
Model (A+B+C+D)	89.67	90.39	89.68	89.93

Table 2: Results for task 2A (disinformation detection).

5.3 Task 2B (Disinformation Class Detection)

In disinformation class detection, as shown in Table 3, Model B achieved a micro-averaged F1-score of 80.36, while Model (B + D) improved performance slightly with a micro-averaged F1-score of 80.71. This shows that combining components B and D enhanced disinformation class detection, emphasizing the value of collaboration between these elements for improved accuracy in identifying specific disinformation classes.

Models	$F1_{mic}$	Pre_{mac}	Rec_{mac}	$F1_{mac}$
Model B	80.36	83.42	80.51	80.36
Model (B + D)	80.71	83.85	80.71	81.81

Table 3: Results for task 2B (disinformation class detection).

In summary, these performance tables demonstrate the power of ensemble models and collaborative approaches in improving the accuracy of persuasion and disinformation detection tasks. Combining different components and models enhanced overall performance, with micro and macro F1-scores consistently rising.

6 Conclusion

In summary, our study emphasizes the power of ensemble models and collaborative approaches in improving the accuracy of persuasion and disinformation detection tasks in Arabic text. We consistently observed enhanced performance through rigorous experimentation, as evidenced by rising micro and macro F1-scores across various model combinations. These results underscore the importance of adaptability and synergy in addressing the nuanced challenges of natural language understanding tasks. Whether it is fine-tuning sentiment semantics, leveraging sentence embeddings, or combining all components, ensemble models consistently outperform individual approaches. These findings offer valuable insights for Arabic text processing and as a model for tackling similar challenges across languages and domains. In an ever-evolving landscape of language processing, our study highlights the significance of diverse techniques and collaborative strategies to effectively meet the complexity of natural language understanding tasks. Ultimately, our research contributes to more accurate solutions and a deeper understanding of persuasive and deceptive language in the digital age.

Limitations

This study, while providing valuable insights into ensemble models for persuasion and disinformation detection in Arabic text, is subject to certain limitations. Using data from limited sources may have restricted the comprehensiveness and generalizability of our findings. Additionally, the complexity and computational demands associated with ensemble models could pose practical constraints in real-world applications, warranting further investigation into model efficiency. Furthermore, the domain-specific focus of our work on persuasion and disinformation detection might limit its direct applicability to other natural language processing tasks or domains. Finally, the interpretability of ensemble models and the potential influence of temporal dynamics in text data represent additional aspects for future research to explore.

Ethics Statement

This research adheres to ethical guidelines and principles in all aspects of data analysis and reporting. The datasets used in this study were sourced from authorized sources, and no personally identifiable information or sensitive data was utilized.

References

Areej Al-Hassan and Hmood Al-Dossari. 2022. Detection of hate speech in arabic tweets using deep learning. *Multimedia systems*, 28(6):1963–1974.

Firoj Alam, Stefano Cresci, Tanmoy Chakraborty, Fabrizio Silvestri, Dimiter Dimitrov, Giovanni Da San Martino, Shaden Shaar, Hamed Firooz, and Preslav Nakov. 2022a. A survey on multimodal disinformation detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6625–6643.

Firoj Alam, Hamdy Mubarak, Wajdi Zaghouani, Giovanni Da San Martino, and Preslav Nakov. 2022b. Overview of the wanlp 2022 shared task on propaganda detection in arabic. *WANLP 2022*, page 108.

Nuha Albadi, Maram Kurdi, and Shivakant Mishra. 2018. Are they our brothers? analysis and detection of religious hate speech in the arabic twittersphere. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 69–76. IEEE.

Abdullah M Alkadri, Abeer Elkorany, and Cherry Ahmed. 2022. Enhancing detection of arabic social spam using data augmentation and machine learning. *Applied Sciences*, 12(22):11388.

Shatha Alyoubi, Manal Kalkatawi, and Felwa Abukhodair. 2023. The detection of fake news in arabic tweets using deep learning. *Applied Sciences*, 13(14):8209.

Alberto Barrón-Cedeno, Israa Jaradat, Giovanni Da San Martino, and Preslav Nakov. 2019. Proppy: Organizing the news based on their propagandistic content. *Information Processing & Management*, 56(5):1849–1864.

Arijit Ghosh Chowdhury, Aniket Didolkar, Ramit Sawhney, and Rajiv Shah. 2019. Arhnet-leveraging community interaction for detection of religious hate speech in arabic. In *Proceedings of the 57th annual meeting of the association for computational linguistics: student research workshop*, pages 273–280.

Giovanni Da San Martino, Yu Seunghak, Alberto Barrón-Cedeno, Rostislav Petrov, Preslav Nakov, et al. 2019. Fine-grained analysis of propaganda in news article. 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 5636–5646. Association for Computational Linguistics.

Anna Feldman, Giovanni Da San Martino, Chris Leberknight, and Preslav Nakov. 2021. Proceedings of the fourth workshop on nlp for internet freedom: Censorship, disinformation, and propaganda. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*.

Khaled M Fouad, Sahar F Sabbeh, and Walaa Medhat. 2022. Arabic fake news detection using deep learning. *Computers, Materials & Continua*, 71(2).

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Ivan Habernal, Raffael Hannemann, Christian Pollak, Christopher Klamm, Patrick Pauli, and Iryna Gurevych. 2017. Argotario: Computational argumentation meets serious games. *EMNLP 2017*, page 7.

Ivan Habernal, Patrick Pauli, and Iryna Gurevych. 2018. Adapting serious game for fallacious argumentation to german: Pitfalls, insights, and best practices. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Fouzi Harrag and Mohamed Khalil Djahli. 2022. Arabic fake news detection: A fact checking based deep learning approach. *Transactions on Asian and Low-Resource Language Information Processing*, 21(4):1–34.

Maram Hasanain, Firoj Alam, Hamdy Mubarak, Samir Abdaljalil, Wajdi Zaghouani, Preslav Nakov, Giovanni Da San Martino, and Abed Alhakim Freihat.

2023. ArAIEval Shared Task: Persuasion Techniques and Disinformation Detection in Arabic Text. In *Proceedings of the First Arabic Natural Language Processing Conference (ArabicNLP 2023)*, Singapore. Association for Computational Linguistics.

Sanaa Kaddoura, Suja A Alex, Maher Itani, Safaa Henno, Asma AlNashash, and D Jude Hemanth. 2023. Arabic spam tweets classification using deep learning. *Neural Computing and Applications*, pages 1–14.

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.

Preslav Nakov, Firoj Alam, Shaden Shaar, Giovanni Da San Martino, and Yifan Zhang. 2021. A second pandemic? analysis of fake news about covid-19 vaccines in qatar. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1010–1021.

Preslav Nakov, Alberto Barrón-Cedeño, Giovanni Da San Martino, Firoj Alam, Julia Maria Struß, Thomas Mandl, Rubén Míguez, Tommaso Caselli, Mucahid Kutlu, Wajdi Zaghouani, et al. 2022. The clef-2022 checkthat! lab on fighting the covid-19 infodemic and fake news detection. In *European Conference on Information Retrieval*, pages 416–428. Springer.

Ali Safaya, Moutasem Abdullatif, and Deniz Yuret. 2020. KUISAIL at SemEval-2020 task 12: BERT-CNN for offensive speech identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2054–2059, Barcelona (online). International Committee for Computational Linguistics.

Mohamad Sharara, Wissam Mohamad, Ralph Tawil, Ralph Chobok, Wolf Assi, and Antonio Tannoury. 2022. Arabert model for propaganda detection. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 520–523.

Zien Sheikh Ali, Watheq Mansour, Fatima Haouari, Maram Hasanain, Tamer Elsayed, and Abdulaziz Al-Ali. 2023. Tahaqqaq: A real-time system for assisting twitter users in arabic claim verification. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3170–3174.

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 Arabic. 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. 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. 2023d. LowResContextQA at Qur'an QA 2023 Shared Task: Temporal and Sequential Representation Augmented Question Answering Span Detection in Arabic. 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-2023 Task 1 2: Enhancing Sentiment Classification and Violence Incitement Detection in Bangla 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)*.