Methodological Expo I — ME–Data Track

(20–25 minutes total)

Key themes:

- **Dataset description**
    - Name, source, size, time range etc.
- **Why this data?**
    - What RQs motivated the use of this data?
        - This should be problem-centered and not data-centered here!
    - What could have been alternatives, and why not those?
    - What makes this data to be unique?
- **Data collection:**
    - Scraping/API/Existing public data released, etc.
    - Permissions/rate limits/access requests
    - Any restrictions?
    - Metadata collected
    - Program/code?
    - Storage formatting
    - Manual steps involved?
    - Show code snippets, flow diagram, collection pipeline here.
- **Challenges faced:**
    - API issues, missing fields, large file handling, computational limitations etc.
    - Bots/spams/noise etc.
- **How did you address the challenges?**
    - Any data cleaning?
    - Any scripts to handle errors?
- **Data cleaning and preprocessing?**
    - Give snapshots of raw data and final data
    - What all was removed and why?
        - E.g., removal of duplicates, missing values handling, removal of non-English content, normalizing timestamps, etc.
- **Reflection:**
    - Ethics of the data collection
    - Potential bias and risks
        - E.g., self-selection, platform demographic skew, etc.
    - Potential misuse of data/ bad actors
    - Reproducibility
        - Can this data be shared broadly?
        - Can the collection and processing scripts be shared publicly?
    - Other interesting comments about the data?

Methodological Expo I — ME–MethodsTrack

(20–25 minutes total)

Key themes:

- **Method description**
    - Name of the method
    - Type of method (statistical / ML / NLP / network / causal / qualitative-computational hybrid, etc.)
    - What problem does this method solve?
    - What type of data does it operate on?
- **Why this method?**
    - What research question motivated the use of this method?
    - Why is this method appropriate for the problem?
    - What could have been alternative methods?
    - Why were those alternatives not chosen?
    - What makes this method particularly useful or powerful here?
        - This should be problem-centered, not method-for-the-sake-of-method.
- **How the method works**
    - Explain clearly and step-by-step:
        - Conceptual intuition behind the method
        - Mathematical or algorithmic logic
        - Inputs required
        - Parameters or hyperparameters
        - Output produced
        - How results are interpreted
    - Show:
        - Workflow diagram/Pseudocode or simplified code snippet/Model architecture (if applicable)/Formula (if statistical method)
        - The goal is clarity, not technical jargon only.
    - **Implementation**
        - What tools/libraries were used?
        - What programming language?
        - Any preprocessing required before applying the method?
        - How long did it take to run?
        - Computational requirements?
        - Any parameter tuning?
        - Show:Code snippet/Pipeline diagram/Example outputs
    - **Challenges faced**
        - Convergence issues, Model instability, Overfitting, Class imbalance, Computational limits, Parameter sensitivity

  - Interpretability, Choosing thresholds
  - Performance evaluation
  - Ground truth limitations
- How did you address the challenges?
  - Cross-validation? Regularization? Alternative specifications? Robustness checks?
- **Evaluation:**
  - How did you assess performance or validity?
  - What metrics were used? (Accuracy, F1, RMSE, AUC, etc.)
  - Any baselines used for comparison?
  - Any robustness checks?
  - Any sensitivity analysis?
- **Reflection:**
  - Limitations and assumptions
  - Bias and fairness concerns
  - Ethical concerns:
    - Risk of misuse
    - Overclaiming causality or predictive power
  - **Reproducibility:**
    - Can code be shared?
    - Are results deterministic?
    - Any randomness or seed sensitivity?
  - **Other interesting comments about the method?**
    - When should someone NOT use this method?
    - How would this scale to larger datasets?
    - Would it generalize to another platform/domain?