Trustworthy and explainable artificial intelligence (AI) lies at the core of building reliable human–AI collaborations. My research investigates how people's behavioral and psychological traits—expressed through online interactions—can inform the design of **fair, interpretable, and emotionally intelligent AI systems**. Despite the rapid progress of large language models (LLMs), their decision-making remains opaque and ethically contested, undermining user trust and limiting adoption. To address this gap, my work bridges social computing, natural language processing (NLP), and trustworthy AI, guided by two overarching questions: **1. How can we derive interpretable representations that reveal users' affective and behavioral attributes on social media?, and 2. How can these insights guide the development of transparent, trustworthy, and ethically aligned AI systems?**

My research answers these questions through a coherent trajectory—from interpretable affect modeling to hybrid explainability, and now toward **Generative Explainable AI (G-XAI)**, a new paradigm that reframes explainability as a generative, adaptive, and user-aware learning process. During my Ph.D., I have authored eleven peer-reviewed publications in venues such as IEEE Access, Springer Nature, and Transactions on Machine Learning Research, including a Best Paper Award at the 2024 IEEE International Conference on Human-Machine Systems (ICHMS). My research has been supported by NSERC, Alberta Innovates, and institutional Transdisciplinary Connector grants, and carried out in collaboration with Alberta Health Services, the Cumming School of Medicine, and political science partners. These interdisciplinary partnerships have bridged AI ethics and governance with healthcare and public policy, grounding my technical contributions in real-world domains. Collectively, this body of work defines my long-term vision: **to build adaptive, generative, and trustworthy AI systems that reason transparently and communicate in ways aligned with human cognition, emotion, and values.**

Interpretable Social-Behavior Analysis

Emotion Detection using Interpretable Handcrafted Features: My early research established the foundation for interpretable affective computing by developing **SSEL** (Figure 1) [Anzum et al., 2023], a domain-specific handcrafted feature representation that captures *Stylistic (S), Sentiment (SE),* and *Linguistic (L)* cues from social-media text. While deep models extract contextual semantics, they obscure causal features and fail to capture personal style. By combining these features through a genetic programming–based ensemble, I demonstrated that SSEL representation provides interpretable insights into the text's emotional context and customization flexibility while exceeding deep learning-level accuracy. This work highlighted that model's interpretability and predictive performance need not be mutually exclusive, setting the methodological basis for my later research on hybrid explainability.

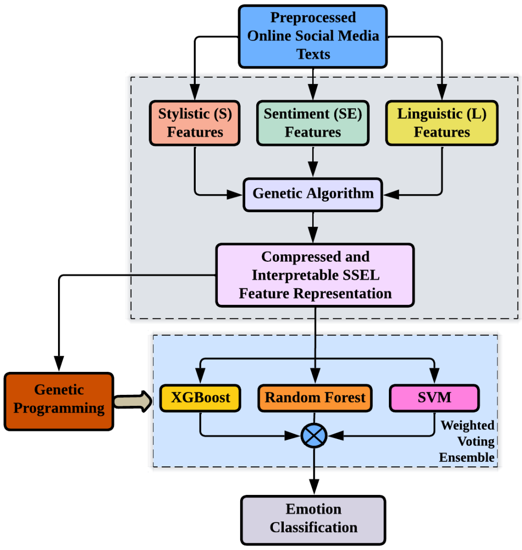

Figure 1: Interpretable SSEL Input Feature Representation for Emotion Detection from Microblogs [Anzum et al., 2023]

EmoBlend Fusion: Toward Trustworthy and Explainable Human–Machine Teaming: Building on SSEL, I developed **EmoBlend Fusion** (Figure 2) [Anzum et al., 2024a], a hybrid deep-learning architecture that integrates handcrafted SSEL features with contextual embeddings from LLMs for emotion detection. The model employs a weighted-ensemble fusion mechanism balancing transparency and performance by integrating interpretable cues with latent semantic representations. A key innovation was a comparative LIME-based explainability analysis between the two modalities, revealing that the handcrafted features captured explicit human-readable justifications for emotion prediction, whereas the LLM embeddings encoded nuanced, context-rich cues, but were less transparent. The findings also showed that these feature spaces to be complementary—offering both interpretability and nuance within a unified trustworthy architecture. Recognized with the **Best Paper Award (ICHMS 2024)** and currently under review at IEEE Transactions on Human-Machine Systems, this work advances explainability toward **trust-calibrated emotion recognition** and sets the stage for adaptive, human-centered AI collaboration.

Trust, Fairness, and Explainability: A Framework for Ethical AI Applications

Mitigating Bias in Social Media Data Mining: While developing interpretable models for affect recognition, I became acutely aware of a deeper challenge: the systemic biases and fairness concerns embedded in social media data and AI-driven predictions. My subsequent research analyzed these biases across the data curation, feature extraction, and model inference stages to understand how emotional and linguistic patterns can inadvertently reproduce demographic or cultural inequalities. I proposed a bias-mitigation pipeline [Anzum et al., 2022] incorporating (1) representational auditing to ensure diversity in datasets, (2) transparent documentation applying Datasheets for Datasets [Gebru et al., 2021] to enhance accountability, and (3) qualitative cross-validation of model outputs to uncover context-dependent distortions. This work reframed bias mitigation as a trust-building process rather than a post-hoc correction, laying the conceptual groundwork for computational models of trust in AI systems.

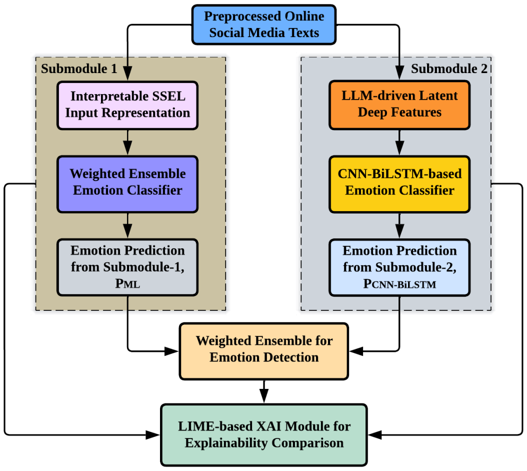
Preprocessed Online Social Media Texts
Submodule 1
Interpretable SSEL Input Representation
Weighted Ensemble Emotion Classifier
Emotion Prediction from Submodule-1, P_{ML}
Submodule 2
LLM-driven Latent Deep Features
CNN-BiLSTM-based Emotion Classifier
Emotion Prediction from Submodule-2, $P_{CNN\text{-}BiLSTM}$
Weighted Ensemble for Emotion Detection
LIME-based XAI Module for Explainability Comparison

Figure 2: EmoBlend Fusion, Incorporating Interpretable SSEL Features with LLM-Driven Latent Deep Features [Anzum et al., 2024a]

Roadmap to Trustworthy AI: Building on my work on user trust in human-centered decision support systems [Gavrilova et al., 2022], I developed a unified framework for trustworthy AI applications [Anzum et al., 2024b] that integrates affective, behavioral and social aspects into a weighted trust computation mechanism. Furthermore, through collaborations with Alberta Health Services, the Cumming School of Medicine, and Political Science, this framework has been employed to evaluate public trust in Canadian healthcare and policy governance, emphasizing three intertwined principles [Dehghani et al., 2026][Dey et al., 2025]:

- **Ethical alignment**: ensuring fairness, privacy, and inclusivity in data-driven pipelines;
- **Explainable processing**: making AI reasoning traceable and interpretable across stakeholders; and
- **Trust calibration**: conceptualizing the adaptive AI systems that adjust behavior and confidence based on user feedback and affective states.

This research positions **trust as an evolving computational relationship**, operationalizing ethical AI through interpretability and behavioral transparency.

Future Directions: Generative, Adaptive and Emotion-Aware Explainability

My next research phase will define the emerging paradigm of **Generative Explainable AI (G-XAI)**—a novel subfield that unites explainability with affective computing and trustworthy interaction. This direction extends my prior contributions into a cohesive new research agenda focused on AI that can dynamically explain itself. Building on these foundations, I will transform explainability from a static, one-size-fits-all afterthought into a dynamic, generative, and socially intelligent process that evolves through human feedback and emotional context.

Emotion-Aware Generative Explanations: Conventional explainability methods optimize fidelity to the model but often ignore whether explanations resonate with the user. I will research the development of **LLM-driven explanation agents** for generating adaptive, context-sensitive explanations that modulate complexity, narrative tone, and modality based on real-time affective signals. Using a fusion of **prompt-conditioned reasoning, af-**

fect recognition, and reinforcement learning, these agents will employ **trust-calibration loops**, where user feedback—confidence ratings, behavioral attributes—serves as a learning signal to refine explanatory style. This approach formalizes explainability as an **interactive meta-learning problem**, where the system continuously updates how it communicates—building sustained, evidence-based trust in long-term human-AI teaming.

Modeling and Quantifying Human Understanding: Despite advances in XAI, it lacks a computational understanding of how users comprehend and internalize AI explanations. To bridge this gap, I plan to develop **cognitive-affective comprehension models** that integrate multimodal signals such as eye-tracking, sentiment evolution, interaction logs, and linguistic embeddings among others to capture how users process and evaluate AI's explanations. These studies will produce **explanation alignment indices** that measure congruence between an AI generated explanation and the user's cognitive and emotional expectations. Such metrics will allow generative models to learn explanation effectiveness directly, advancing explainability from a one-way transparency mechanism to a **bidirectional, human–AI dialogue** characterized by mutual intelligibility.

Ethical and Societal Alignment: To ensure these adaptive explanation techniques are deployed responsibly, I will embed them within a broader trustworthy AI pipeline that includes bias auditing, fairness-constrained optimization, and differential privacy safeguards. Through transdisciplinary collaborations in domains like healthcare, education, and online safety, my lab will develop auditable end-to-end pipelines to trace explanation provenance, quantify biases in generated justifications, and ensure equitable communication across diverse user groups. This work will produce **open-source evaluation toolkits and policy guidelines** for emotion-aware generative AI, setting new standards for equitable, trustworthy human-AI collaboration.

Vision and Impact: Over the next five years, I will establish the **Trustworthy Human–AI Interaction Lab**, uniting generative modeling, affective computing, and cognitive science to design AI systems that sense, explain, and adapt. Drawing on my experience with cross-disciplinary initiatives, I will pursue funding from national and international agencies and cultivate partnerships with industry stakeholders committed to trustworthy and responsible AI. Within the G-XAI paradigm, my group will pioneer **Generative, Adaptive and Emotion-Aware Explainability** as a new frontier of AI research, integrating affective understanding with generative reasoning to create empathetic and cognitively aligned AI explanations. In this vision, explainability becomes an *active, ongoing dialogue*—AI systems that justify, listen, and recalibrate — transforming trust from a passive belief into an active, co-constructed understanding between humans and intelligent systems.

References

[Anzum et al., 2024a] **Fahim Anzum** and Marina L. Gavrilova, "EmoBlend Fusion: Leveraging Handcrafted and Deep Features for Emotion Detection," in *IEEE 4th International Conference on Human-Machine Systems **(ICHMS)***, 2024.

[Anzum et al., 2024b] **Fahim Anzum**, Ashratuz Zavin Asha, Lily Dey, Artemy Gavrilov, Fariha Iffath, Abu Quwsar Ohi, Liam Pond, MD Shopon, and Marina L. Gavrilova, A Comprehensive Review of Trustworthy, Ethical, and Explainable Computer Vision Advancements in Online Social Media, in *Global Perspectives on the Applications of Computer Vision in Cybersecurity*, IGI Global, 2024.

[Dehghani et al., 2026] Farzaneh Dehghani, Mahsa Dibaji, **Fahim Anzum**, Lily Dey, Alican Basdemir, Sayeh Bayat, Jean-Christophe Boucher et al., "Trustworthy and Responsible AI for Human-Centric Autonomous Decision-Making Systems.", accepted to be published in *Transactions on Machine Learning Research (TMLR)*, 2026.

[Anzum et al., 2023] **Fahim Anzum** and M. L. Gavrilova, "Emotion Detection From Micro-Blogs Using Novel Input Representation," in *IEEE Access*, 2023.

[Anzum et al., 2022] **Fahim Anzum**, A. Z. Asha and M. L. Gavrilova, "Biases, Fairness, and Implications of Using AI in Social Media Data Mining," in *IEEE 21st International Conference on Cyberworlds **(CW)***, 2022.

[Gavrilova et al, 2022] Gavrilova, M.L., **Fahim Anzum**, and 6 others, A Multifaceted Role of Biometrics in Online Security, Privacy, and Trustworthy Decision Making. In *Breakthroughs in Digital Biometrics and Forensics.* Springer, 2022.

[Gebru et al, 2021] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. "Datasheets for datasets." in *Communications of the ACM*, 2021.

[Dey et al., 2025] Lily Dey, **Fahim Anzum**, Ulises Charles-Rodriguez, A S M Hossain bari, Jean-Christophe Boucher, Aleem Bharwani, Marina L. Gavrilova, "Exploring Public Trust Through LLM-Driven Opinion Mining", in *International Conference on Computer Information Systems and Industrial Management (CISIM)*, Cham: Springer Nature, 2025.