Recent advancements in transformer-based models have enabled researchers to use large language models (LLMs) to analyze human behavioral and psychological traits on social media, providing insights into users' needs and preferences. Despite these capabilities, LLMs' inherent complexity and unresolved ethical concerns undermine user trust and limit the adoption of frameworks that integrate human data. This raises two key questions central to my research: **1. How can we extract interpretable features to capture users' behavioral attributes on social media?**, and **2. What ethical and trust considerations are essential for creating a unified trustworthy AI application framework?** My research focuses on developing interpretable methods to capture affective and linguistic patterns, integrating them with LLMs and deep learning (DL) for social media opinion mining. I have led a cross-disciplinary project that applied this approach to analyze Canadians' trust perceptions in health, providing insights for medical institutions on key health determinants. At the Biometric Technologies Lab, I have spearheaded the development of methodologies for a trustworthy and explainable AI (TXAI) framework, incorporating diverse user trust factors to enhance transparency, reliability, and alignment in AI-driven decisions.

During my Ph.D., I have published **eight papers** in reputed computer science and transdisciplinary venues, including *IEEE Access, Springer Nature, IEEE Cyberworlds, and IEEE Human-Machine Systems*. My recent work on a bi-modal architecture for interpretable emotion detection earned the **Best Paper Award** at the 2024 IEEE International Conference on Human-Machine Systems **(ICHMS)**, highlighting its contribution to advancing *Trustworthy Human-Machine Teaming for Effective Decision-Making*. As a researcher dedicated to advancing fair and trustworthy AI, I have successfully secured provincial and federal funding from **Alberta Innovates**. My ongoing Ph.D. project, funded by **NSERC, Alberta Innovates,** and **Transdisciplinary Connector Grants**, addresses ethical AI challenges in social media data mining, emphasizing trust, bias, fairness, and explainability.

Interpretable Social-Behavior Analysis

Emotion Detection using Interpretable Handcrafted Features: My initial research focused on developing handcrafted features to capture domain-specific emotional cues in social media text. While DL models offer rich contextual insights, their black-box nature limits interpretability and overlooks user-specific stylistic patterns. I proposed a novel interpretable handcrafted feature representation employing a genetic algorithm to combine three user-specific features: **Stylistic (S), Sentiment (SE), and Linguistic (L)** (Figure 1) [Anzum et al., 2023]. The combined **SSEL** representation was used to train a weighted voting ensemble of ML models, fine-tuned through genetic programming. I demonstrated that SSEL representation provides interpretable insights into the text's emotional context, customization flexibility, and improved accuracy. This research underscores **the importance of incorporating interpretable feature engineering techniques in predictive systems using human data, where ensuring transparency and interpretability is crucial.**

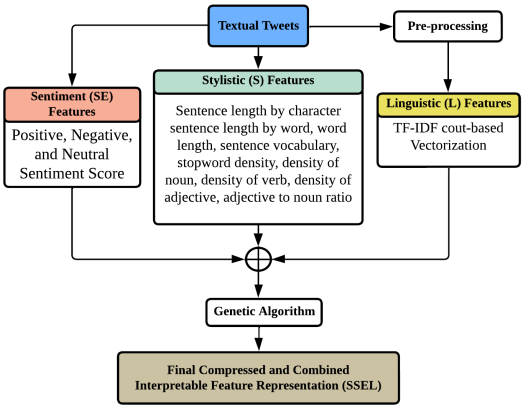

Figure 1: I developed SSEL to interpret user behavior for emotion detection.

EmoBlend Fusion: Promoting Trustworthy Human-Machine Teaming: My recent research introduced **EmoBlend Fusion**, a hybrid model that integrates SSEL's interpretable features with rich contextual LLM-derived deep features (Figure 2) [Anzum et al., 2024a]. By employing a weighted ensemble mechanism to fuse the predictions of the SSEL feature-based module with the LLM-driven feature-based modules, this bi-modal architecture utilizes the complementary strengths of diverse feature groups, improving accuracy, robustness, and interpretability. To my knowledge, this is **the first work to combine interpretable handcrafted features with LLM-driven contextual deep features** for bi-modal emotion detection, laying the groundwork for fostering robust and trustworthy human-machine teaming by capturing behavioral cues with greater precision.

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

Mitigating Bias in Social Media Data Mining: My research explores the ethical AI challenges in conventional social media data mining approaches, focusing on trust, bias, fairness, and explainability [Anzum et al., 2022]. I addressed three underexplored questions: What biases are embedded in social media data mining? How do AI models address (or fail to address) fairness in their predictions? And what risks arise from leveraging social media data in AI workflows? This analysis highlighted how biases in data preparation, processing, and modeling stages distort AI predictions, leading to unfair outcomes and eroding users' trust and privacy protections. I outlined bias mitigation strategies: 1) Curating diverse datasets to better represent the heterogeneity of user demographics and experiences, 2) Implementing datasheets for datasets [Gebru et al., 2021] to enhance transparency and accountability, and 3) Incorporating qualitative data analysis to uncover and address nuanced biases. Together, these efforts address technical and systemic ethical challenges in social media data mining, emphasizing fairness and transparency. By integrating practical guidelines with actionable insights, this research advocates for user-centered AI practices, urging the community to prioritize fairness, privacy, and trust.

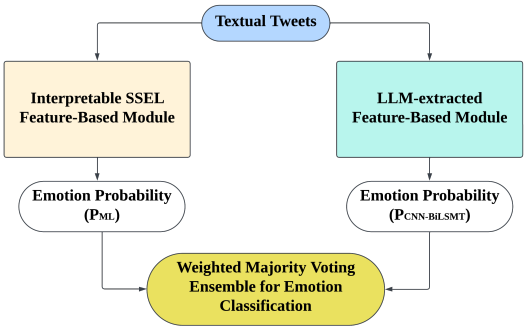

Figure 2: EmoBlend Fusion, incorporating the interpretable features with deep features.

Roadmap to Trustworthy AI: Building on previous research 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]. Central to this framework is a novel trust computation model incorporating affective traits, personality, and social influences. I introduced a weighted trust assessment mechanism integrating user-specific factors into the evaluation process. The framework emphasizes ethical data practices, robust processing, bias mitigation, and explainability to tackle challenges in social media applications. I collaborated with interdisciplinary teams to explore ethical AI principles, focusing on governance, regulatory compliance, and reproducibility [Dehghani et al., 2024]. This research emphasized the importance of transparent communication throughout the AI lifecycle and improved bias detection through multi-source data analysis, cross-model evaluation, and iterative testing. The work produced actionable guidelines for aligning AI technologies with societal values, supporting the responsible deployment of AI across various domains.

Future Research

I aim to spur further the development and evaluation of **adaptive, context-aware, and human-centered AI (HCAI) systems that identify human needs while ensuring explainability, fairness, and trustworthiness in AI solutions.** I plan to focus on the following research areas in the next 3-5 years.

Personalization through Behavioral Insights: I explored how personality traits and emotions influence decision-making and preferences across diverse contexts. However, existing systems, such as personality and affect-aware recommender systems, fail to capture the dynamic nature of these traits and limit their ability to provide adaptive recommendations. More research is needed to understand how these traits impact users' perceptions of recommendation quality. Moreover, current approaches overlook the fluctuating interplay between affective states and external factors, highlighting the need for real-time adaptive models. To address these gaps, my research will develop hybrid approaches that combine behavioral science insights with advanced AI techniques to create systems that adapt to users' evolving emotional and personality traits. These models will incorporate real-time context and user feedback to provide personalized decision support. Through empirical studies and user-centered evaluations, I will also investigate how personality traits and emotions influence users' perceptions of recommendation quality. Additionally, I will explore how affective states interact with external factors to improve the accuracy of preference prediction in dynamic environments. I will focus on privacy-preserving mechanisms and transparent frameworks for consent and data use to ensure ethical integrity. Ultimately, I aim to advance personalized AI by designing adaptive, user-centered, and ethically responsible systems.

Writing User-Centered Explanation using LLM: While developing frameworks to analyze Canadians' trust perceptions in health systems [Dey et al., 2024], I identified the need for contextual explanations to enhance trust and support informed decision-making. I plan to investigate how generative AI and LLMs can produce transparent, personalized, and contextually adaptive explanations that evolve with users' preferences and behavioral traits. I will explore how users from diverse demographics and cultural and linguistic backgrounds perceive these explanations' relevance, clarity, and trustworthiness. Through qualitative studies and user evaluations, I aim to address the limitations of current XAI methods that either overwhelm users with technical complexity or provide overly simplistic explanations. My broader vision is to advance TXAI, focusing on the impact of explanations in fostering trust, informed decision-making, and equitable human-machine interactions.

Advancing Fairness in AI through Cross-Disciplinary Collaboration: Building on my expertise in trustworthy AI and ethical decision-making, my future research will focus on creating innovative frameworks to mitigate bias in healthcare predictive models, addressing critical issues of social justice and accessibility. By systematically exploring the relationship between sensitive attributes, fairness metrics, and bias mitigation strategies, I will focus on ensuring that AI systems are equitable and transparent and serve marginalized populations who face systemic disparities in healthcare. Committed to advancing cross-disciplinary research, I plan to collaborate with health informatics, public policy, social computing, and human-centered AI researchers to develop transformative tools for equitable healthcare decision-making. This work will enhance fairness in AI while driving tangible improvements in healthcare access, equity, and outcomes, ensuring that technological innovation leads to meaningful societal impact.

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., 2024] 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.", 2024, *Preprint.*

[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., 2024] Lily Dey, **Fahim Anzum**, Ulises Charles-Rodriguez, A S M Hossain bari, Jean-Christophe Boucher, Aleem Bharwani, Marina L. Gavrilova, "A Digital Lighthouse: Exploring Health Concerns and Public Trust Using LLM-Driven Opinion Mining from Canadian Reddit Communities", under review in *PLOS One*, 2024.