# Universidade de Pernambuco
# Programa de Pós-Graduação em Engenharia da Computação (PPGEC)

## Proposta de Dissertação de Mestrado

**Área: Computação Inteligente ou Modelagem Computacional**

**Título: Modelagem Computacional da Comunicação Não-Violenta com Apoio de *Large Language Models* e Análise Contínua de Emoções**

**Orientador: Sérgio Murilo Maciel Fernandes (sergio.fernandes@upe.br)**

**Descrição** - A intensificação das interações digitais tem ampliado a incidência de linguagem ofensiva e padrões comunicacionais agressivos, impulsionando pesquisas em Processamento de Linguagem Natural voltadas à detecção de discurso de ódio e toxicidade textual. Revisões sistemáticas evidenciam a complexidade conceitual do problema, marcada por múltiplas definições, diferentes níveis de agressividade e desafios metodológicos associados à anotação e avaliação de dados textuais (FORTUNA; NUNES, 2018; SCHMIDT; WIEGAND, 2017). Evidências empíricas indicam, ainda, que manifestações ofensivas frequentemente emergem de construções linguísticas sutis e contextuais, associadas a vieses sociais e semânticos, e não apenas a símbolos explícitos (TALAT; HOVY, 2016).

No campo da análise afetiva, abordagens clássicas de *opinion mining* estabeleceram fundamentos sólidos, porém limitados pela adoção de categorias discretas de sentimento (PANG; LEE, 2008). Em contraste, modelos contemporâneos defendem a representação contínua das emoções, permitindo a quantificação de estados afetivos por meio de dimensões psicológicas como Valência, Excitação e Dominância. A formulação da análise emocional como um problema de regressão tem se mostrado particularmente adequada para capturar nuances emocionais e estimar graus de agressividade de forma mais precisa, apoiando-se em léxicos afetivos com validação empírica rigorosa (BUECHEL; HAHN, 2017; MOHAMMAD, 2018).

Paralelamente, os *Large Language Models* (LLMs) redefiniram o estado da arte em tarefas de linguagem natural ao demonstrar capacidades avançadas de generalização, aprendizado em contexto e interpretação semântica profunda (BROWN *et al*., 2020). Levantamentos recentes indicam seu potencial para aplicações sensíveis, incluindo a análise e mitigação de conteúdos tóxicos, embora desafios relacionados à avaliação, confiabilidade e controle de comportamento permaneçam em aberto (ZHAO *et al*., 2023). Resultados recentes apontam, inclusive, a viabilidade do uso direto de LLMs na detecção de toxicidade textual, aproximando modelos de grande escala de demandas sociais associadas à comunicação ética (KUMAR *et al*., 2023).

Nesse contexto, esta proposta visa investigar um arcabouço computacional para apoio à comunicação não-violenta, integrando métricas estatísticas de linguagem ofensiva, modelagem contínua de emoções e LLMs como camadas semânticas de alto nível. Modelos de aprendizado eficientes, como *Extreme Learning Machines*, são considerados como alternativas metodológicas para tarefas de regressão emocional sob restrições computacionais, em função de sua simplicidade arquitetural e baixo custo de treinamento (HUANG *et al*., 2006). O trabalho busca contribuir metodologicamente ao explorar a integração entre diferentes paradigmas de modelagem e, empiricamente, ao avaliar sua eficácia em cenários de linguagem ofensiva e emocionalmente carregada. Como desdobramento, a pesquisa estabelece bases conceituais para extensões futuras envolvendo análise e correção multimodal de sentimentos (BALTRUSAITIS *et al*., 2019).

**Referências:**

BALTRUŠAITIS, Tadas; AHUJA, Chaitanya; MORENCY, Louis-Philippe. Multimodal machine learning: A survey and taxonomy. **IEEE transactions on pattern analysis and machine intelligence**, v. 41, n. 2, p. 423-443, 2018.

BROWN, Tom et al. Language models are few-shot learners. **Advances in neural information processing systems**, v. 33, p. 1877-1901, 2020.

BUECHEL, Sven; HAHN, Udo. Emotion analysis as a regression problem—dimensional models and their implications on emotion representation and metrical evaluation. In: **Proceedings of the Twenty-second European Conference on Artificial Intelligence**. 2016. p. 1114-1122.

FORTUNA, Paula; NUNES, Sérgio. A survey on automatic detection of hate speech in text. **Acm Computing Surveys**, v. 51, n. 4, p. 1-30, 2018.

HUANG, Guang-Bin; ZHU, Qin-Yu; SIEW, Chee-Kheong. Extreme learning machine: theory and applications. **Neurocomputing**, v. 70, n. 1-3, p. 489-501, 2006.

KUMAR, Deepak; ABUHASHEM, Yousef Anees; DURUMERIC, Zakir. Watch your language: Investigating content moderation with large language models. In: **Proceedings of the International AAAI Conference on Web and Social Media**. 2024. p. 865-878.

MOHAMMAD, Saif. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In: **Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)**. 2018. p. 174-184.

PANG, Bo; LEE, Lillian. Opinion mining and sentiment analysis. **Foundations and Trends® in information retrieval**, v. 2, n. 1–2, p. 1-135, 2008.

SCHMIDT, Anna; WIEGAND, Michael. A survey on hate speech detection using natural language processing. In: **Proceedings of the fifth international workshop on natural language processing for social media**. 2017. p. 1-10.

TALAT, Zeerak; HOVY, Dirk. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In: **Proceedings of the NAACL student research workshop**. 2016. p. 88-93.

ZHAO, Wayne Xin et al. A survey of large language models. **arXiv preprint arXiv:2303.18223**, v. 1, n. 2, 2023.