Research Proposal for Ph. D. Application processing

by, Mr. Aravind Chandran, M.S.

Application to: Department of Computer Science and Engineering, School of Computing

Research area: Natural Language Processing, Machine Unlearning and Machine Learning

Mode of Research: Part-time

Proposal of work: In today's real-world dynamic environments, NLP models are deployed to a greater extent in applications like chatbots, search engines, educational platforms, etc., their ability to learn continuously from new data is crucial. However, this versatility comes with growing concerns about what these models retain. In real-world settings, models may internalize and perpetuate social biases, toxic language, or misinformation present in their training data. Moreover, continual learning can exacerbate this issue by accumulating undesirable knowledge over time, i.e., they must not only learn continuously but also forget selectively. This capability – machine unlearning – the ability to selectively remove harmful or outdated information from language models by exploring methods that allow models to "forget" specific patterns (e.g., toxic, biased, or misleading content) while preserving useful knowledge and overall performance. The goal is to enable ethical, flexible, and regulation-aware NLP systems that remain trustworthy over time.

Based on the above scenario, the central motivation would be: **How to design an efficient and scalable unlearning mechanism for NLP models that selectively forget biased, toxic, or misleading language patterns in a continual learning framework?**

This study is timely and impactful given the growing concern over AI-generated misinformation, embedded social biases, and the legal need for data erasure (e.g., under General Data Protection Regulation - GDPR). It bridges ethical AI, machine learning, and NLP, aiming to make language models safer and more adaptable.

The research may involve:

- Constructing annotated datasets for toxic language, biased patterns, and misinformation.
- Implementing continual learning NLP models that support modular training and forgetting.
- Designing unlearning algorithms that target specific harmful content while preserving general performance.
- Evaluating forgetting efficacy, retention of unrelated knowledge, and mitigation of downstream harms using tasks like sentiment analysis, question answering, and natural language inference.