PROJECT DESCRIPTION:

Natural Language Processing (NLP) has been front and center with the increasing focus on Large Language Models (LLM) such as ChatGPT. Sentiment analysis of online product reviews is an important problem in NLP and LLM. With knowledge gained from sentiment analysis, business analysts can better their products as well as improve targeted advertisement.

This project focuses on machine learning models for sentiment analysis using product reviews. While the dataset contains products from multiple online companies, for this project, Amazon will be the target retailer. The data subset contains about 5000 records. The research team will perform the following key tasks:

(1) Identify two approaches to text mining using machine learning for sentiment analysis. The machine learners of interest are XGBoost, CatBoost, Naïve Bayes, Random Forest, and Neural Networks.

(2) Develop and analyze machine learning models using the Amazon reviews subset consisting of about 5000 records. In addition to evaluating individual learners, an Ensemble Learner of the three best models will be performed. All empirical development will be done in the Python for machine learning environment.

(3) The dataset is imbalanced, thus, research into best practices for class imbalance mitigation will be performed, and consequent data sampling experiments will be incorporated into Step (2).

(4) Draft a technical report detailing all the steps and processes of this research project. A conference and/or journal paper for publication will be developed, post completion of the project and technical report; however, drafting the for-publication paper is not a deliverable of the project.

Audience are academic and industry practitioners of sentiment analysis within the domain of NLP and LLM. The student team will gain valuable knowledge in text mining, data sampling, machine learning, and data imbalance problem resolution. Oral and writing presentation are learned / improved as soft skills.

Deliverables	Type of work	Activities	Resources	Tech Skills	Priority
Literature Review that leads to a specific research question agreed with the client.	Review for minimal two sentiment analysis approaches, as verified in literature. Techniques to explore, e.g.: (1) N-gram text2vector based (2) Emoticons based (3) SentiWordNet based Client will choose two of these three.	Read research articles related to NLP and sentiment analysis, and briefly summarize their work and contributions. The articles to read will be provided by the client.	Library access to research articles from conferences and journals across various publishing venues.	Mathematics. Statistics. Python Basics. Databases. Reading. Writing.	high
Low-fidelity solution(s) – preliminary machine learning models.	Generate ideas and create solutions with high level of content but low specificity based on the findings of literature. Exploring the five machine learners. Could be far from the final machine learning models developed.	Brainstorming sessions, regular online meetings with project mentor. Brief presentations of to-date research as planned under the project.	Faculty mentorship. Python programming language and libraries. Online resources related to NLP and sentiment analysis and data sampling	Mathematics. Statistics. Python. Machine Learning Basics in Python, Databases. Communication. Online research.	mid
Analysis of the dataset. Machine learning models for sentiment analysis. Python program scripts. Research notebook.	Data sampling (preprocessing), Modeling and analysis of the sentiment analysis using five machine learners, Ensemble-based modeling and analysis, Python-based ML solutions, Continued development of research notebook.	Data analysis and preprocessing. Machine learning modeling and analysis. Comparative study of at least five machine learners built with dataset. Ensemble-based modeling and analysis.	Python libraries for machine learning, primarily scikit-learn. Python script development for the respective machine learners.	Python programming language. Modeling and Analysis. Machine Learning. Databases. Mathematics. Communication.	high
Technical report of research work. Written presentation of research work.	Drafting a technical report that provides a detailed description of the empirical studies performed under the project.	Documenting all the activities and results of the research work in the journal paper. May be accompanied by a detailed enough technical report to help with the publication preparation.	Project notebook, Python programs, and Other information maintained throughout the research project. Faculty mentorship.	Machine Learning. Technical Report Writing. Mathematics. Report's Draft Proofing. Communication.	high
Oral presentation of research work.	An oral presentation (pptx format) made to project team and other audiences.	Preparing effective written and visual presentation of research project.	MS Word and PowerPoint. LaTeX may also be used for publication drafting.	Oral presentation. Writing skills. Visual presentation.	high