CSE674: Text Analytics

Credit Hours Structure	(3,0,3) = (Theory, Lab, Total)
Prerequisites	Machine Learning-I
Co-requisites	None
Barred Combinations	None
Course Lead / E-mail	Sajjad Haider (sahaider@iba.edu.pk)
Last Modified	January 12, 2024

Course Description

The course delves into the fundamentals of text analytics, which involve applying natural language processing and machine learning techniques to derive valuable insights from unstructured text data. It covers essential aspects of text data mining and analysis, while also exploring recent advancements in deep learning applied to natural language processing (NLP). Topics include encoder-decoder architecture, attention and self-attention mechanisms, word embeddings, transformers, and language models.

The course goes in-depth into large language models (LLMs), including their pre-training and fine-tuning processes, as well as their diverse applications across various NLP tasks. Students will gain hands-on experience with transformer-based language models using the Hugging Face ecosystem.

Throughout the course, students actively participate in projects and assignments that emphasize the practical implementation of text analytics techniques. They also critically review and present influential research papers in the field, fostering a comprehensive understanding of both historical developments and state-of-the-art technologies in NLP.

Learning Outcomes

By the end of this course, students are expected to:

- Develop a strong foundational understanding of key text analytics techniques, including text categorization, sentiment analysis, document clustering, summarization, topic modeling, feature engineering, and word embeddings.
- Gain expertise in recent advancements in deep learning for Natural Language Processing (NLP), including encoder-decoder architecture, attention and self-attention mechanisms, and transformer models.
- Acquire in-depth knowledge of large language models, their pre-training, fine-tuning, and extensive application across a broad spectrum of NLP tasks.
- Develop practical skills in working with the Hugging Face ecosystem, a leading platform for transformer-based models.

(Tentative) Weekly Plan

Week	Topics
1	Overview of course, Challenges in text analytics, Introduction to HuggingFace based Transformer models for Sentiment Analysis, NER, Question Answering, Summarization, Machine Translation and Text Generation, SQUAD Dataset, Green vs Red AI, Cost of Training LLM
2	Bag of Word model, Term-Document representation, Feature Extraction, Uni-grams, bi-grams and n-grams, TF and TF-IDF, Similarity Measures
3	Applications of ML Algorithms (Decision Tree, Random Forest and Naïve Bayes) for Text Classification and Sentiment Analysis, Classification using Pre-trained Models, Assignment 1 Issued
4	Explainable/Interpretable AI (LIME), Lexicon-based Sentiment Analysis, WordNet, Introduction to Feedforward Neural Networks, Word Embeddings, Word2Vec, CBOW and Skip-gram Architectures
5	Embeddings (Cont'd): Subsampling and Negative Sampling, Doc2Vec, GloVe and Other Embeddings, Contextual Embeddings, Vector DB
6	Semantic Text Similarity, Retrieval Augmented Generation (RAG), Document Clustering and Evaluation, Assignment 1 Due
7	Project Topic Proposal Discussion, Recap of Linear Algebra, Singular Value Decomposition, Latent Semantic Anaysis (LSA), Queries in LSA,
	Midterm Exam and Midterm Break
8	Text Summarization: Graphs based Algorithm, LSA based and TextRank Algorithms, Summarization Evaluation: ROUGE, GLUE, Assignment 2 Issued
9	Overview of Neural Networks, Recurrent Neural Networks (RNN), Long Short Term Memory (LSTM), Seq2Seq Model, Encoder-Decoder Architecture
10	Attention and Self-Attention Architecture, Transformer, Markov Model and Text Generation
11	Introduction to Large Language Models, Open Sourced and Quantized Versions, Prompt Engineering, Assignment 2 Due
12	Fine-tuning LLMs, Supervised and Semi-Supervised Learning Methods, Parameter Efficient Fine Tuning (PEFT), Low-Rank Adaption Methods (LoRA and QLoRA)
13	Issues in Building LLM from Scratch, Evaluation Benchmarks: ARC, MMLU, Hellaswag, TruthfulQA, Recent Trends in NLP and LLM, Course Recap
14	Project Presentation

(Tentative) Marks Distribution

- 2 Assignments – 16%
- Project - 20%
- Paper Presentation and Class Participation – 8%
- Midterm – 26%
- Final - 30%

Note: *Assignments, projects, and paper presentations will be done in groups to encourage discussion and collaboration within groups.*

Assignment 1: Text Classification and Sentiment Analysis: Develop a text classification system that categorizes text documents into predefined classes and performs sentiment analysis on a dataset of reviews or social media posts.

Assignment 2: Text Summarization: Implement a text summarization tool offering extractive and abstractive methods and compare their performance.

Potential Projects: Although groups will have a chance to suggest their own project, and after approval, they can work on the suggested topic. However, considering the popularity of LLM these days, one of the following two areas could be chosen for the projects.

1. **Retrieval Augmented Generation (RAG) Projects:** Develop a RAG-based question answering system that retrieves information from a given dataset (like Wikipedia or a specific domain corpus) to answer questions posed in natural language.
2. **Low-Rank Adaptation Methods (LoRA) Projects:** Use LoRA to adapt a pre-trained language model for a specific domain (e.g., legal, medical, or technical text).

Seminal Paper Presentation: Throughout the semester, each group will present one of the seminal papers in NLP (such as the one on BERT, BART, GPT, GPT-2, etc.) or the new breakthroughs (Mamba, Mixture of Experts, etc.). Each presentation should cover the paper's key contributions, methodologies, results, and impact on the field. A 10-15 minute presentation will be followed by a Q&A session.

Reference Readings/Books:

Being a rapidly evolving field, the course would make use of recent papers and blogs (primarily hosted on medium.com), but the following books will be utilized for certain topics:

- Natural Language Processing with Transformers by Tunstall et al. (2022)
- Transformers for Natural Language Processing by Rothman (2022)
- Practical Natural Language Processing by Vajjala et al. (2020)
- Text Mining with Machine Learning by Zizka et al. (2020)
- Hands-on Python Natural Language Processing by Kedia and Rasu (2020)
- Practical Text Analytics by Anandarajan, Hill and Nolan (2019)
- Natural Language Processing in Action by Lane et al. (2019)