FA25-DATA-0297 Course Fall 2025

Introduction to Natural Language Processing

Instructor & Staff

- Shuo Zhang, Ph.D., Professor of the Practice, Data Analytics, Tufts University (Office Hours: Monday 4-5 pm, JCC 684 or by appointment (zoom))
- Jeanie Cherng, Ph.D, Industry Advisor
- Ishmam Khan, Research Assistant

Email

Shuo.Zhang@tufts.edu

Class Time/location:

Mondays 6-9pm, Tisch Library 316, Medford/Somerville campus

Course GitHub Repo

github repo link

Course Website (schedule)

Please refer to the course website for the updated schedule.

1 Syllabus

1.1 Pre-requisite

Proficient in python coding; familiarity with probability, linear algebra, and calculus, and machine learning (intro level)

1.2 Course Description

NLP is now at the center of AI, data science and data analytics. There is a wealth of textual data online. NLP-enabled products constitute an essential part of everyday life, both in consumer facing products (Siri, Alexa, ChatGPT, Google Translate, etc.) and B2B applications (e.g., NLP for medical and legal domains). However, understanding human languages and extracting structured information from this plethora of unstructured text data is a major challenge for modern computers. The recent advancement in machine learning and deep learning makes NLP one of the fastest growing fields in AI and data science. In this course, we will survey ML-based NLP techniques from statistical ML approaches to the SOTA deep learning models. Students will be able to apply the ML and NLP skills learned in this course to real world problems in a variety of jobs (NLP/ML engineer, NLP data scientist, data scientist, data analyst, etc) and industries/sectors (finance, medical insurance, technology, urban planning, education, etc.).

1.3 Textbooks

We will use a combination of online reading materials and Jupyter Notebook tutorials, as well as online versions of the updated textbooks in deep learning and NLP.

- *Speech and Language Processing*, 3rd edition, Jurafsky and Martin.

1.4 Grades

Course grades will be a weighted sum of two factors: homework and coding assignments (60%), and a final project (40%).

1.5 Coding Assignments

Coding assignments will be using python and the tensorflow keras API. To complete a coding assignment, first code out the assignment yourself in a Jupyter notebook. Work in your branch and commit your code locally in Git. You are encouraged to meet with your teammates for a discussion before pushing your final code. For details see the course website "assignment and presentation work flow".

1.6 Final Project

Students will be divided into 1-3 person teams to propose a final project to apply NLP techniques to a real-world problem or dataset. Students will use the final four weeks of the course to conduct independent research to formulate the project and give a final presentation. The final project is evaluated by the final presentation and a final paper. In the final project, each student in the team must contribute at least one method that has not been covered in the lectures and compare them with simple baselines. Please see final project rules for details.

1.7 Academic Policies & Student Support

Academic Integrity Policy: Tufts holds its students strictly accountable for adherence to academic integrity. Student Resources: Religious Accommodations: Tufts University faculty, staff, and administration highly value and acknowledge the religious diversity of its student body. Accommodations for Students with Disabilities: Tufts University values the diversity of our students, staff, and faculty and recognizes the important contribution each student makes to our unique community.

2 Specifics

2.1 Homework & code reviews

One team will be responsible for delivering a code review for a homework on the due date (Monday in class). I will post the schedule of which team to present which homework. The code review could feature one or more presenters.

2.2 Reading discussions

Each team will be responsible for discussing and presenting one of the readings in class. All team members should get together to discuss the reading in your team meeting prior to the class, and then one or more team members will present the reading in class and lead a discussion.

2.3 Final Project rules

2.3.1 General

Find a partner to form a group and define a final project and carry out a month long research and implementation. Group size can vary from 1-3. The final project will be delivered in a presentation and a paper. You should produce a paper that you can aim to submit to a conference for publication (although one month is definitely not enough for that, but that's the direction of the goal we are trying to achieve here).

There are several types of final project:

- solve a well defined NLP task using a dataset and compare several different methods (training models from scratch or fine tuning LLM pretrained models)

- using prompt engineering for LLM to solve a NLP task
- Location Extraction project (NER) from Prof Alice Mello
- exploratory analysis of text data (see NLP4DH below)
- replicate a research paper
- propose a new method on a task
- etc.

No matter the type of project, in each project:

- You must explore the use of a novel method that was not covered in this class, and this is something that you should do research on and learn about yourself. This could also be exploring a new task that was not covered in detail in class and explore methods to be applied to that task.
- The co-authors must each contribute at least one novel method to be used in this project and final paper
- The paper should compare several methods/approaches to a simple baseline and report back the results.
- You must have a limited scope to be able to finish within about 1 month.
- Time estimation. For instance, how long do you plan to do literature review, data collection and curation, writing code, etc. This is just to make sure you have a realistic chance of finishing a project in a month. For instance, if the problem you want to solve involves a very long process of data collection, you might need to think about whether you have time to finish it.
- Think about how you will evaluate your methods. If you need annotated ground truth data, think about how you will obtain it.

2.3.2 Project proposal

Components of the project (proposal) :

- Dataset
- Data preprocessing/cleaning
- Research question
- Problem description
- Scope
- Literature review (1–2 papers per person)
- Methodology (compare several novel methods not covered in class to a simple baseline)
- Evaluation
- Error analysis
- Timeline estimation

2.3.3 Deliverables

The final project will be delivered in the form of a presentation and a paper, not code.

2.3.4 Proposal Requirements

In your proposal, you can cover some of the above aspects, but it is not necessary to include all of them at the proposal stage. The proposal should primarily focus on:

- Research question
- Problem description and motivation
- Dataset / data acquisition plan
- Scope and planned literature review

2.3.5 Grading (Total: 40 Points)

- **Paper formatting and style** (5 points)
  The main paper content (excluding references) should be a minimum of 4 pages and a maximum of 6 pages, in a two-column format. The paper should be formatted using the LaTeX or Word template available here:
  `https://2023.aclweb.org/calls/style_and_formatting/`
- **Result & discussion** (5 points)
  Is there a thorough discussion on results and observations, error analysis, etc.?
- **In-class presentation** (5 points)
  Does the in-class presentation clearly discuss the motivation, problem formulation, dataset, and proposed methods?
- **Literature review** (5 points)
  Thoroughly review and explain relevant existing work in this domain.
- **Novelty and Difficulty** (5 points)
  How novel is this idea? How challenging is this project to execute? This is correlated with the amount of work invested. For easier problems, more effort is expected (e.g., experimenting with more advanced methods). For harder problems, lower performance is acceptable if justified.
- **Experimental Validation** (5 points)
  Is there sufficient experimental validation to support the main claim or conclusion?
- **Clarity of Writing** (5 points)
  Is the paper clearly written? Can a reader easily understand what was done? Be specific and transparent about your work.
- **Technical Correctness** (5 points)
  Is the technical solution correct?

3 Course schedule

Please refer to the course website.