The University of Texas at Dallas
Applied Natural Language Processing
Spring 2024
BUAN 6342

**Course Information**

| | |
|---|---|
| Course Number/Section: | BUAN6342.S01 |
| Course Title: | Applied Natural Language Processing |
| Day & Times: | Monday 10:00am -12:45pm |
| Classroom: | JSOM 1.502 |

**Instructor Contact Information**



| | |
|---|---|
| *Professor* | Harpreet Singh |
| *Office Phone* | 972-883-4770 |
| *Email Address* | harpreet@utdallas.edu |
| *Office Location* | JSOM 3.430 |
| *Office Hours:* | Monday 4pm to 6pm |
| *TA Office Hours:* | By Appointment |



**Course Pre-requisites**

Prerequisites: BUAN 6341

**Course Description**

This is an advanced course focusing on natural language processing and the utility of textual data to gain meaningful quantitative and actionable insights about the language (mainly English) using rule-based and statistical methods and to extract the information for real-world applications. This class will focus both on modern neural methods for these problems.

**Course Learning Objectives**

By the end of this course, students will be able to

- Apply existing libraries (including scikit-learn, gensim, spacy, PyTorch, HuggingFace) to text data.
- Use these tools for classification problems (sentiment analysis, spam detection etc.), Machine translation, Name Entity Recognition, Summarization, Question Answering, and Semantic Search.

**Course Access and Navigation**

The course can be accessed using your UT Dallas NetID account on the eLearning website. Please see the course access and navigation section of the Getting Started with eLearning webpage for more information.

**Communication**
We will use regular email (harpreet@utdallas.edu) and a web conferencing tool (Teams) during the semester. Please copy TA on all communications. **Student emails and messages will be answered within 3 working days under normal circumstances**.

**Textbook:**
There is no required Textbook for the course.

**Recommended Textbooks:**
The following texts are useful, but they are optional.

- Lewis Tunstall, Leandro von Werra, Thomas Wolf Natural Language Processing with Transformers,
- Dan Jurafsky and James H. Martin. Speech and Language Processing (3rd ed. draft) **– Free**
- Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep Learning **– Free**

**Tools and Languages used**
- **Language** : Python
- **Deep Learning Framework**: Pytorch (mostly), HiggingFace
- **Python Packages**: Gensim, Spacy, NLTK, Scikit-Learn, etc.

**Class Lectures**

This course will be taught in a traditional in-class mode. The lectures will be a mixture of presentations, code explanations, and code-along sessions. The instructor might post some videos as well.

**Class Materials**

The Instructor may provide class materials that will be made available to all students registered for this class as they are intended to supplement the classroom experience. These materials may be downloaded during the course; however, these materials are for registered students' use only. Classroom materials may not be reproduced or shared with those not in class, or uploaded to other online environments except to implement an approved Office of Student Access Ability accommodation. Failure to comply with these University requirements is a violation of the Student Code of Conduct.

**Homework Assignments:**
The Homework assignments **require individual work**. You will submit your assignments in the required file format through eLearning. Instructions and dates for each assignment will be posted on eLearning.

**Late work Policy for HWs and Project:**

- Homework assignments will be turned in on eLearning and are due at 11:59pm CT on their assigned due date.
- **Each student has a total of three free late days to use on assignments throughout the semester**.
- There is no restriction on how you can use the late days. You can use the late days on one assignment or across multiple assignments.

- You do not need permission to use free late days. We will keep track of late days that you have used.
- After you have exhausted free late days, you will receive a penalty of 25% per late day. You do not need to take any permissions. The minimum score in the assignment will be 0.

**Exams:**
There are two exams for this course.
Both exam will be proctored at the UTD Testing Center. **Students will need to make reservation using "Reserve Your Seat" tool at least 72 hours prior to the exam time**. Please see the UTD Testing Center web pages https://ets.utdallas.edu/testing-center/ for more information. Please be sure to review the Testing Center Student Guidelines.

---

**Grading Policy**

Your grade in the course will be based on the following items:

| | |
|---|---|
| Class Participation | 10 |
| Homework Assignments | 50 |
| Exam 1 | 20 |
| Exam 2 | 20 |
| Total | 100 |

These are the only scores that will be used to determine your grade. **There will be no extra work or repeat exams will be given and no late work will be accepted, so please do not ask for an exception**. Your semester average will be rounded using the convention in rounding and your final letter grade will be determined as follows:

**Grading Scale:**
Relative grading allows educators to convert the outcomes of a student's test, project or assignment and adjust that final grade in relation to grades from other students in the course. Relative grading is similar to bell curving or grading on a curve, and considers the highest score as the baseline (A), relatively adjusting all others compared to that score. Student should earn a passing grade for each projects and exam grading components in order to be considered for a letter grade in the range of C to A. Note: this grading system is following the UTD/JSOM policy to keep the class grade average between B to A-.

Your course grade will depend on your overall score relative to your peers.

1) The students with scores in the 80th percentile and above will get an A grade.
2) The students with scores between the $80^{th}$ and the $60^{th}$ percentile will get an A- grade.
3) The students with scores between the $60^{th}$ and the $40^{th}$ percentile will get a B+ grade.
4) The students with scores between the $40^{th}$ and the $20^{th}$ percentile will get a B grade.
5) The students with scores between the $20^{th}$ and the $10^{th}$ percentile will get a B- grade.
6) The instructor will decide the students with scores below the 10th percentile.

## Tentative Class Schedule

***These descriptions and timelines are subject to change at the discretion of the instructor.***

**Additional notes, instructions and useful links will be posted on eLearning/Teams,** so please make sure to stay current on materials and announcements posted between meetings. ***You are responsible for keeping up with all posted material and announcements, so make sure you check eLearning/Teams on a daily basis.***

| Date | Content |
|---|---|
| 15th Jan | ○ Martin Luther King Holiday |
| 22nd Jan | ○ Intro Course<br>○ Spacy and Pre-Processing<br>○ Sparse Embeddings (tfidf) |
| 29th Jan | ○ Sentiment Analysis with Spacy and sklearn<br>○ Into Neural Networks -I (Theory) |
| 5th Feb | ○ Into Neural Networks -II (Theory)<br>○ PyTorch Layers<br>○ Numeric Prediction with PyTorch<br>○ Classification with PyTorch<br>○ Introduction to Embeddings<br>○ IMDB Sentiment Analysis with simple DL (PyTorch) |
| 12th Feb | ○ Intro to Hugging Face<br>○ IMDB Sentiment Analysis with simple DL (PyTorch + Hugging Face)<br>○ Word2Vec/Language Models/RNN/LSTM (Theory) |
| 19th Feb | ○ Encoder-Decoder/Attention/Transformer (Theory)<br>○ Understanding BERT/GPT/T5 (Theory)<br>○ IMDB Sentiment Analysis using BERT<br>○ Revisit Sentiment analysis (Domain Adaptation – BERT) |
| 26h Feb | ○ Finding duplicates in a given a pair of Quora questions (Sentence pair classification using BERT<br>○ Stack exchange multi-label classification using BERT<br>○ Revisit IMDB sentiment Analysis (Handling Longer Sequences)<br>○ Named Entity Recognition (Token Classification) |
| 26th Feb | ○ Kaggle Competition |
| 4th march | ○ Current Models (GPT4, LLAMA, Mistral)<br>○ Parameter Efficient fine tuning (LLAMA)<br>○ OpenAI Finetuning |
| 11th march | ○ Break |
| 18th – 22nd March | ○ Exam1 |
| 25th March | ○ Intro to Sentence-BERT (Sentence Embeddings)<br>○ Finding top-k similar questions<br>○ BerTopic (Topic Modelling using Bert) |
| 1st April | ○ Few to No Label (Zero Shot classification, data augmentation) |
| 8th April | ○ Revisit the Language model (focus on inference)<br>○ Beam Search/Greedy Decoding<br>○ Sampling in Text Generation (Top-k, Top-p, Temperature, Greedy, Random)<br>○ Sequence to Sequence Tasks |

| | |
|---|---|
| | ○ Translation/Summarization/Simple Question-Answer |
| 15th April | ○ Intro to Prompt Engg (PE), Semantic Search (SS), RAG, LangChain (LC), Vector Database (VD) -I<br>○ Kaggle Competition Report – Mid report/Mandatory First Submission |
| 22nd April | ○ PE, SS, RAG, LG, VD -II |
| 29th April | ○ PE, SS, RAG, LG, VD -III |
| 6th May – 10th May | Exam2 |
| 10th May | Final Kaggle Competition Report/ No more submissions allowed |

**Make-Up policy**

You are required to take all homework assignments, exams and projects on the designated dates and with your own class. Generally, there will be no makeup quizzes or exams given. If you believe you cannot take an exam or quiz on the regularly scheduled date, you should talk to the instructor as soon as possible. If you cannot take an exam or quiz due to a compelling personal reason such as emergency surgery or death in the immediate family, you have to notify me before the exam. Supporting documentation, such as hospital admission, will be required. Routine or regular doctor's office visits will not be an acceptable excuse. Failure to give notification before the exam will result in an automatic 15% deduction for the quiz/exam grade if a makeup is approved. There will be no make-up of assignment if not submitted by due date.

**Academic Integrity**

The faculty and administration of the School of Management expect from our students a high level of responsibility and academic honesty. Because the value of an academic degree depends upon the absolute integrity of the work done by the student for that degree, it is imperative that a student demonstrate a high standard of individual honor in his or her scholastic work. We want to establish a reputation for the honorable behavior of our graduates, which extends throughout their careers. Both your individual reputation and the school's reputation matter to your success.

Dishonesty includes, but is not limited to plagiarism, cheating, collusion, facilitating academic dishonesty, fabrication, failure to contribute to a collaborative project and sabotage.

- **Plagiarism:** The adoption or reproduction of ideas, words, statements, images or works of another person as one's own without proper acknowledgement.
- **Cheating:** Using or attempting to use unauthorized materials, information, or study aids in any academic exercise. Academic exercise includes all forms of work submitted for credit or hours.
- **Fabrication:** Falsification or creation of any information, data or citation in an academic exercise.
- **Collaboration and/or Collusion:** Seeking or providing aid to another student in completion of any assignment submitted for academic credit without permission from the faculty member. http://www.utdallas.edu/judicialaffairs/UTDJudicialAffairs-Basicexamples.html

Students in this course suspected of academic dishonesty are subject to disciplinary proceedings, and if found responsible, the following minimum sanctions will be applied:

- **Assignment – Zero for the Assignment**
- **Exams – F for the course**

These sanctions will be administered only after a student has been found officially responsible for academic dishonesty, either through waiving their right for a disciplinary hearing, or being declared responsible after a hearing administered by Judicial Affairs and the Dean of Student's Office.

In the event that the student receives a failing grade for the course for academic dishonesty, the student is not allowed to withdraw as a way of preventing the grade from being entered on their record. Where a student receives an F in a course and chooses to take the course over to improve their grade, the original grade of F remains on their transcript, but does not count towards calculation of their GPA.

The School of Management also reserves the right to review a student's disciplinary record, on file with the Dean of Students, as one of the criteria for determining a student's eligibility for a scholarship.

**UT Dallas Syllabus Policies and Procedures**

The information contained in the following link constitutes the University's policies and procedures segment of the course syllabus. Please go to http://go.utdallas.edu/syllabus-policies for these policies.