Course Syllabus, Version 10.12.2020

GRAD-E1347: Natural Language Processing with Deep Learning

Concentration : Policy Analysis

Slava Jankin and Hannah Bechara

1. General information

Class time	Tue, 10-12h
Course Format	This course is taught online only via the platform Clickmeeting/Teams. Clickmeeting/Teams allows for interactive, participatory, seminar style teaching.
Instructor	Slava Jankin and Hannah Béchara
Instructor's office	3.15 and 3.14
Instructor's e-mail	jankin@hertie-school.org, bechara@hertie-school.org
Instructor's phone number	Slava Jankin: +49 30 259 219 167 Hannah Béchara: +49 30 259 219 252
Assistant	Name: Alex Karras Email: karras@hertie-school.org Phone: +49 30 259 219 156 Room: 3.45
Instructor's Office Hours	Upon request

Link to Module Handbook MIA and MPP
Link to Study, Examination and Admission Rules

Instructor Information:
Slava Jankin is Professor of Data Science and Public Policy at the Hertie School. He is the Director of the Hertie School Data Science Lab. His research and teaching is primarily in the field of natural language processing and machine learning. Before joining the Hertie School faculty, he was a Professor of Public Policy and Data Science at University of Essex, holding a joint appointment in the Institute for Analytics and Data Science and Department of Government. At Essex, Slava served as a Chief Scientific Adviser to Essex County Council, focusing on artificial intelligence and data science in public services. He previously worked at University College London and London School of Economics. Slava holds a PhD in Political Science from Trinity College Dublin.

Hannah Béchara is an NLP post-doc who inadvertently found herself hired by Hertie's Data Science Lab. In between training neural networks and support vector machines, Hannah occasionally teaches programming classes in Python, the programming language for winners. She has previously been spotted teaching classes on NLP methods and Maths for Machine Learning. Hannah's current research interests include semantic relationships between words and phrases, and encompasses entailment, contradictions, and causal relations. Most importantly, Hannah plans to use NLP to solve all of the world's problems. For reasons yet unclear, the University of Wolverhampton decided to award Hannah a PhD in Computer Science.

2. Course Contents and Learning Objectives

Course contents:

Natural Language Processing (NLP) is a key technology of the information age. Automatically processing natural language outputs is a key component of artificial intelligence. Applications of NLP are everywhere because people and institutions largely communicate in language. Recently statistical techniques based on neural networks have achieved a number of remarkable successes in natural language processing leading to a great deal of commercial and academic interest in the field. This course provides an overview of modern data-driven models to richer structural representations of how words interact to create meaning. We will discuss salient linguistic phenomena and successful computational models. We will also cover machine learning techniques relevant to natural language processing.

Main learning objectives:

In this course, students will gain a thorough introduction to cutting-edge research in Deep Learning for NLP. Through lectures, assignments and a final project, students will learn the necessary skills to design, implement, and understand their own neural network models.

Target group:

Students interested in developing strong methodological foundations for machine learning research and practice.

Teaching style:

Lectures covering theoretical concepts followed by practical lab sessions. This is an intensive course with a significant research component undertaken by the students.

Prerequisites:

Python Programming (E1326).

Software:

We will be using production-ready Python frameworks like PyTorch. In addition, for practical work we will make heavy use of Jupyter notebooks, Google Colab, and GitHub.

Diversity Statement:

As you may know, the Hertie School is committed to implementing a new Diversity and Inclusion Strategy. We strive to have an inclusive classroom but ask your informal feedback on inclusivity throughout the course.

3. Grading and Assignments

Composition of Final Grade:

Assignment 1: **Project Proposal and Literature Review**	Deadline: Session 4	Submit via Moodle	20%
Assignment 2: **Midterm Report**	Deadline: Session 7	Submit via Moodle	20%
Assignment 3: **Final Report**	Deadline: Session 11	Submit via Moodle	40%
Assignment 4: **Presentation**	Project Presentations: Session 12	Submit via Moodle	10%
Participation grade			10%

The assessment for the course consists of a research project, presentation and participation. The research project must be done in teams of 2-4 (individual submissions will not be accepted for the project). The aim is to develop research projects as close as possible to an academic publication in the area of applied machine learning and communicate your research to the broader public.

The aim of the assessments is three-fold:

- First, it will provide you with the opportunity to apply the concepts learned in this class creatively, which helps you with understanding material more deeply.
- Second, designing and working on a unique project in a team which is something that you will encounter, if you haven't already, in the workplace, and the project helps you prepare for that.
- Third, along with the opportunity to practice and the satisfaction of working creatively, students can use this project to enhance their portfolio or resume. We will discuss with individual project groups whether they can be turned into academic publications

Note about grading. There is no "perfect project." While you are encouraged to be ambitious, the most important aspect of this research project is your learning experience. Hence, you don't want to pick something that is too easy for you, but similarly, you don't want to choose a project where you are not certain that is out of the scope of this class. The project proposal is not graded by how exciting your project is but based on whether you follow the objectives of the project proposal, project presentation, and project report. For instance, if your project ends up being unsuccessful – for example, if you choose to design a classifier and it doesn't achieve the desired accuracy – it will not negatively affect your grade as long as you are honest, describe the potential issues well, and suggest improvements or further experiments. Again, the objective of this project is to provide you with hands-on practice and an opportunity to learn.

Assignment Details

Assignment 1: Project proposal and literature review (20%) – 3 pages and 5 references

- The main purpose of the project proposal is to receive feedback from the instructor regarding whether your project is feasible and whether it is within the scope of this class. Also, the project proposal offers a chance to receive useful feedback and suggestions on your project. The goal is for you to propose the research question to be examined, motivate its rationale as an interesting question worth asking, and assess its potential to contribute new knowledge by situating it within related literature in the scientific community.

- For the project, you will be working in a team consisting of 2-4 students. The members of each team will be randomly assigned by the instructor. If you have any concerns about working with someone in your group, please discuss it with the instructor.
- You must include a link to a GitHub repository containing the code of your project. Your repository must be viewable to the instructor by the submission deadline. If your repository is private, make it accessible to us (GitHub IDs *sjankin* and *hbechara*). If your repository is not visible to us, your assignment will not be considered complete, so if you are worried please submit well in advance of the deadline so we can confirm the repository is visible. Furthermore, we will assess individual contribution to the team, should such an issue arise, based on the frequency and quality of GitHub commits in your project repository, so make sure you start the repository as the very first stage of your project.
- After you have received feedback from the instructor and your project proposal has been graded, you are advised to stick to the project outline in the proposal as closely as possible. However, if there is a concept introduced in a later lecture, you have the option to modify your proposal, but you are not penalized if you don't. If you wish to update your project outline, talk to the instructor first.
- The LaTeX template for the proposal and detailed description of the content and the marking rubric will be made available on Moodle.

Assignment 2: Midterm report (20%) – 4 pages and 10 references

- By the middle of the course, students should present initial experimental results and establish a validation strategy to be performed at the end of experimentation. This serves as a project milestone. The milestone should help you make progress on your project, practice your technical writing skills, and receive feedback on both.
- Ultimately, your final report will be written in the same style as an NLP research paper. For the midterm, we ask you to write a preliminary version of some sections of your final report. Producing a high-quality milestone is time well-spent, because it will make it easier for you to write your final report. You might find that you can reuse parts of your project proposal in your milestone. This is fine, though make sure to act on any feedback you received on your proposal.
- The LaTeX template for the proposal and detailed description of the content and the marking rubric will be made available on Moodle.

Assignment 3: Final report (40%) – 8 pages and unlimited references

- The final report will include a complete description of work undertaken for the project, including data collection, development of methods, experimental details (complete enough for replication), comparison with past work, and a thorough analysis. Projects will be evaluated according to standards for conference publication—including clarity, originality, soundness, substance, evaluation, meaningful comparison, and impact (of ideas, software, and/or datasets).
- You ***must*** include a link to a GitHub repository containing full replication code of your project.
- The LaTeX template for the proposal and detailed description of the content and the marking rubric will be made available on Moodle.

Assignment 4: Presentation (10%)

- At the end of the semester, teams will produce a blogpost (use this template: https://github.com/hertie-data-science-lab/distill-template) and pre-recorded video presenting the results of their work to the class and broader community. These will be posted on the Data Science Lab website.
- Detailed description of the presentation task will be made available on Moodle.

Participation grade (10%)

- We appreciate everyone being actively involved in the class. For full participation credit, we expect you to contribute relevant questions and ideas to the online class forum on Piazza, and answer questions from your peers. The top ~5 contributors will get full participation grade; others will get credit in proportion to the contribution of the ~5th person. Use your real name and your Hertie email address for participation credit. We will regularly show the "leaderboard" of contributors in class.

  Any other act that improves the class for this year or subsequent years, which the instructor notices and deems worthy will receive 1% credit.

Late submission of assignments: For each day the assignment is turned in late, the grade will be reduced by 10% (e.g. submission two days after the deadline would result in 20% grade deduction).

Attendance: Students are expected to be present and prepared for every class session. Active participation during lectures and seminar discussions is essential. If unavoidable circumstances arise which prevent attendance or preparation, the instructor should be advised by email with as much advance notice as possible. Please note that students cannot miss more than two out of 12 course sessions. For further information please consult the Examination Rules §10.

Academic Integrity: The Hertie School is committed to the standards of good academic and ethical conduct. Any violation of these standards shall be subject to disciplinary action. Plagiarism, deceitful actions as well as free-riding in group work are not tolerated. See Examination Rules §16.

Compensation for Disadvantages: If a student furnishes evidence that he or she is not able to take an examination as required in whole or in part due to disability or permanent illness, the Examination Committee may upon written request approve learning accommodation(s). In this respect, the submission of adequate certificates may be required. See Examination Rules §14.

Extenuating circumstances: An extension can be granted due to extenuating circumstances (i.e., for reasons like illness, personal loss or hardship, or caring duties). In such cases, please contact the course instructors and the Examination Office *in advance* of the deadline.

4. General Readings

- Jeremy Howard, Sylvain Gugger. 2020. *Deep Learning for Coders with fastai and PyTorch*. O'Reilly Media, Inc. [we'll designate it as **HG** throughout]
- Aston Zhang and Zachary C. Lipton and Mu Li and Alexander J. Smola. 2020. *Dive into Deep Learning*. URL: https://d2l.ai. [we'll designate it as **ZLLS** throughout]
- Dan Jurafsky and James H. Martin. *Speech and Language Processing.* 3rd edition draft. https://web.stanford.edu/~jurafsky/slp3/ [we'll designate it as **JM** throughout]

5. Session Overview

Session	Session Date	Session Title
1	09.02.2021	**Deep Learning Journey**
2	16.02.2021	**From Model to Production**

3	23.02.2021	Data Ethics
4	02.03.2021	Preliminaries
5	09.03.2021	Linear Neural Networks
6	16.03.2021	Multilayer Perceptrons
Mid-term Exam Week: 22 – 26.03.2021 – no class		
7	30.03.2021	Recurrent Neural Networks
8	06.04.2021	Modern Recurrent Neural Networks
9	13.04.2021	Attention Mechanisms
10	20.04.2021	Pretraining
11	27.04.2021	Applications
12	04.05.2021	Project Presentations
Final Exam Week: 17 – 21.05.2021 – no class		

6. Course Sessions and Readings

In the case that there is a change in readings, students will be notified by email.

Required readings are to be read and analysed thoroughly. Optional readings are intended to broaden your knowledge in the respective area, and it is highly recommended to at least skim them.

Lab sessions will use textbook notebooks: https://github.com/ageron/handson-ml2

Session 1: Deep Learning Journey	
Learning Objective	We introduce key concepts behind deep learning. Try out first deep learning models.
Required Readings	HG: Chapter 1 ZLLS: Chapter 1
Optional Readings	
Lab content	https://github.com/fastai/fastbook/blob/master/01_intro.ipynb

Session 2: From Model to Production	
Learning Objective	We discuss the best practices of using deep learning in practice. We cover end-to-end process of creating a deep learning application. We discuss capabilities and constraints of deep learning and key learning points for practical deployment.
Required Readings	HG: Chapter 2

Optional Readings	
Lab content	https://github.com/fastai/fastbook/blob/master/02_production.ipynb

Session 3: Data Ethics	
Learning Objective	What happens when things go wrong? Or your deep learning model does things that it shouldn't? We will cover this core questions in practical applications of deep learning. In the lab, we will also introduce HuggingFace Transformers.
Required Readings	HG: Chapter 3
Optional Readings	
Lab content	https://huggingface.co/transformers/quicktour.html

Session 4: Preliminaries	
Learning Objective	We recap key concepts for storing, manipulating, and pre-processing data. We also revisit key topics in linear algebra, calculus, and probability.
Required Readings	ZLLS: Chapter 2
Optional Readings	
Lab content	ZLLS: Chapter 2

Session 5: Linear Neural Networks	
Learning Objective	We introduce the simplest neural networks and cover the training process.
Required Readings	ZLLS: Chapter 3
Optional Readings	
Lab content	ZLLS: Chapter 3

Session 6: Multilayer Perceptrons	
Learning Objective	We introduce deep networks and focus on the issues of overfitting, underfitting, and model selection.
Required Readings	ZLLS: Chapter 4
Optional Readings	
Lab content	ZLLS: Chapter 4

Mid-term Exam Week: 22 – 26.03.2021 – no class

Session 7: Recurrent Neural Networks	
Learning Objective	We introduce architectures that are designed to handle textual data.
Required Readings	ZLLS: Chapter 8
Optional Readings	
Lab content	• ZLLS: Chapter 8 • https://github.com/fastai/fastbook/blob/master/10_nlp.ipynb

Session 8: Modern Recurrent Neural Networks	
Learning Objective	We introduce more sophisticated architectures that are designed to deal with numerical instability of RNNs. We cover GRUs, LSTMs, and other architectures.
Required Readings	ZLLS: Chapter 9
Optional Readings	
Lab content	• ZLLS: Chapter 9 • https://github.com/fastai/fastbook/blob/master/12_nlp_dive.ipynb

Session 9: Attention Mechanisms	
Learning Objective	We cover some of the most powerful recent ideas in NLP that use attention and transformers.
Required Readings	ZLLS: Chapter 10
Optional Readings	
Lab content	ZLLS: Chapter 10

Session 10: Pretraining	
Learning Objective	We cover key pretraining models in NLP from embeddings to BERT.
Required Readings	ZLLS: Chapter 14
Optional Readings	
Lab content	ZLLS: Chapter 14

Session 11: Applications

Learning Objective	We bring together previously discussed techniques into full NLP pipelines and introduce two specific tasks: sentiment analysis and natural language inference.
Required Readings	ZLLS: Chapter 15
Optional Readings	
Lab content	ZLLS: Chapter 15

Session 12: Project Presentations

Final Exam Week: 17 – 21.05.2021 – no class