Hertie School

Master Thesis Supervision Plan
Academic Year 2020-2021
MIA/MPP

Name:	Slava Jankin
Professor of…	Data Science and Public Policy

Advisor Information

Office	3.15
E-Mail	jankin@hertie-school.org
Faculty Assistant	Alex Karras
Office Hours	tba

Research & Advising Profile:
General Academic Field: What is your academic field, and what are your areas of supervision?

My general academic field is data science. My research and teaching are primarily in the fields of natural language processing (NLP) and machine learning (ML). Generally, I will supervise theses that focus on applications of data science, ML and NLP techniques in substantive research or methodological work.

Specific Expertise: What is your specific area of expertise, and what topics are you supervising?

Data science, ML and NLP applications in climate change and health, international development, public administration, and comparative politics.

Methodology: What methodologies and methods are you able to supervise?

Machine learning and natural language processing

Supervision Style: Please include information on what your advisees can expect from you in terms of supervision format (for instance, are you supervising by colloquium only; are you offering individual meetings; are you supervising team projects?) Please be informed that there will be both onsite/online options according to your availability:

Supervision will be conducted by colloquia and, if needed, individual meetings. Team projects are also welcome. The goal is to develop good projects into publishable scientific papers.

Examples of Previously Supervised Thesis Topics: (*if applicable*)

"Mapping Poverty in Bangladesh with Satellite Images and Deep Learing" by Dang Ngoc Huy (Class of 2020)

Proposed Projects: (*if applicable*)

Project Title	Practice Partner (if applicable)
There is significant debate and discussion in the literature on approaches to international development. Yet, we know little about how different countries have approached the issue of international development, and how this has changed over time. We know little about the key differences in how richer countries and poorer countries approach international development. This project will look at how "international development" is discussed in the annual UN General Debates and the evolution of the term over time. In particular, it will aim to identify the key development issues raised by different countries, and how these issues have changed over time. It will aim to identify the international development issues that have widespread support among states in the UN, and those that are most contested.	
A significant literature looks at the influence of international aid on foreign policy preferences. See: http://aiddata.org/replication-datasets. In general, these studies have tended to use voting in the UN General Assembly to examine whether aid is used by donor governments to "buy" votes. There has been little consideration of whether aid can lead to a deeper shift in foreign policy preferences due to changing the norms etc in the recipient government. This is surprising given the literature on aid and democratisation suggests that aid can have this positive effect. This project will draw on the UN General Debate Corpus to examine whether foreign aid can lead to such a deeper shift. The focus is on assessing whether an aid relationship between two countries leads to a convergence of foreign policy preferences. In addition, the project will consider an interaction of aid (and budget support) and democratisation that may result in shifts in foreign policy preferences.	
The transition from the Millennium Development Goals (MDGs) to the Sustainable Development Goals (SDGs) brought with it significant changes in the process of creating the goals and with the actual content of the SDGs. One of the most important developments was the inclusion of SDG 16, which recognises the central role of effective, accountable and inclusive political institutions in promoting sustainable development. The interconnectedness of the 17 SDGs has been widely discussed in the literature. Many, however, argue that progress on SDG 16 in particular, is crucial for achieving the other SDGs. This is because the Goal is centred on the organisation of power in society and nature of governance, which is crucial for implementing policies that positively impact issues such as poverty, hunger, health, education, gender equality, water and sanitation, and climate change. Without good governance and strong institutions, it will not be possible to address these issues captured in the other SDGs. Yet, a significant shortcoming is the difficulty in measuring progress on this SDG 16. In addition to general issues linked with data availability across the various indicators, a key challenge is aggregating trends across these wide-ranging indicators to track overall progress on SDG 16. A second issue that follows, is that despite claims regarding the centrality of SDG 16 for achieving the other SDGs,	

little is known about the causal pathways from the different indicators in SDG 16 to the other SDGs and associated indicators. This project will explore how changes in SDG 16 indicators impact a country's progress towards indicators linked to health, gender equality, water and sanitation, and climate change. The aim is to use data science and machine learning techniques to track progress on SDG 16 by bringing together data across the different indicators; and measure the impact of the different aspects of SDG 16 on other SDGs.	

Plagiarism:

Plagiarism is an infringement of § 11 Good Academic Conduct, 2a: "Infringements of the standards of good academic conduct include for instance to use wordings, ideas or other intellectual work of others in an academic work without clearly indicating the author. The obligation to indicate the authorship of others shall apply irrespective of whether or not the sources used are protected by copyright" (See: Exam Rules, § 11 Good Academic Conduct for more information).

It is vital to keep track of your sources and to cite all material properly.

The Library will offer a session available to all students on resource management and proper citation.

Extra (individual) note on plagiarism

Attendance:

Students receive 8 ECTS for the Master´s thesis colloquium. Attendance for supervision—whether in the form of a colloquium or an individual meeting—is compulsory. Students are expected to be present, prepared, and engaged in each session, and to adhere to deadlines set for assignments.

In case of an excused absence, students must notify their supervisor. Absence must be compensated with an additional assignment.

Students on an academic exchange programme in the Fall Semester should be given the opportunity of remote supervision.

Participation & Milestones:

1. 3-page project proposal
2. 3-page Related Work section
3. 5-page Proposed Method, Experiments, and Future Work sections
4. 2-page Analysis section
5. 15-page (without references and appendix) final version
6. Digital poster mock up

Colloquium & Meeting Information:

	Colloquium Dates	Session Title
Meeting times (suggested)	Session 1: 3 November 2020	Scoping and planning
	Session 2: 24 November 2020	Research context
	Session 3: 2 February 2021	Midterm milestone
	Session 4: 16 February 2021	Analysis
	Session 5: 30 March 2021	Final version
	Session 6: 13 April 2021	Last check and project defence

Colloquium Sessions:

Session 1: 03.11.2019 **Scoping and planning**	
Aim	**Discuss the proposal and develop a timeline for the project**
Assignment	You need to draft a 3-page project proposal and send it to me by **8am, 3 November 2020**. The proposal should contain the following sections: 1. Discuss ***one*** key academic paper that is the most relevant to your project. You should summarise the paper, providing the title of the paper, the goal and achievement of the paper, your criteria for selecting the paper, solution proposed in the paper, data used in the paper, methods used in the paper, references to other (up to) 5 relevant papers. 2. Motivation: why your project is interesting; describe the main goals of the project; what is the scientific question you are trying to answer; why you chose this goal 3. Task: what will you try to do in the thesis 4. Data: what data will you use 5. Method: what method(s) you are planning to use 6. What baseline will you choose 7. How would successful outcome of your project look like 8. Timeline of the project, including a Gannt chart. You should use Overleaf to draft the proposal in LaTeX. By the first meeting you should also set up the GitHub repository for the project. It can remain "private" at this stage, but I should have access to it. Overleaf also integrates with GitHub!
Readings (if applicable)	

Session 2: 24.11.2020

Research context	
Aim	Understand the research context of the project by providing an overview of existing work in the area
Assignment	You need to draft a 3-page Related Work section of your thesis (on Overleaf) and send it to me by **8am, 24 November 2020**. The section should discuss papers that inspired your approach, papers that you use as baselines, papers proposing alternative approaches to the problem, papers applying your methods to different tasks, etc. You shouldn't go deep into detail on any single paper, instead you should explain how the papers relate to each other and how they relate to your work. You should take note of the datasets used in the papers, methods developed and used, evaluation strategies, and theoretical arguments. Google Scholar will be come your best friend at this stage. Remember that it has advanced search options!
Readings (if applicable)	

Session 3: 02.02.2021 **Midterm Milestone**	
Aim	You should present initial experimental results and establish a validation strategy to be performed at the end of experimentation. At the very minimum, you should show that you have setup your data, baseline model code and evaluation metric, and run experiments to obtain some results. This stage should help you make progress on your thesis, practice your technical writing skills, and receive feedback on both.
Assignment	You need to draft three sections (5 pages) on Overleaf: - Proposed method: details your approaches to the problem (e.g. architecture of the model and any other key methods or algorithms). You should be specific and detailed in describing your main approaches, you may want to include equations and figures. You should also describe your baseline(s). You should clearly state whether your approaches are original or provide necessary references. - Experiments: describe the dataset(s) used (provide references); discuss the evaluation metric(s) you used and any other details needed to understand your evaluation strategy; discuss how you ran the experiments, providing necessary technical details; and comment on your qualitative results. - Future work: describe what you plan to do for the rest of the project and why. You need to send it to me by **8am, 2 February 2021**. Your GitHub should also reflect all the work done at this stage.

Readings (if applicable)	

Session 4: 16.02.2021 **Analysis**	
Aim	Discuss final results and qualitative analysis of the findings.
Assignment (e.g. first chapter)	You should update your experimental results if you carried out additional experiments. You will need to add a 2-page Analysis section to your Overleaf draft. You need to send it to me by **8am, 16 February 2021**. The section will contain qualitative evaluation of the results. You should try to understand your system (how it works, when it succeeds and when it fails) by measuring or inspecting key characteristics or outputs of your model. You could comment on selected examples, develop error analysis, measure the performance metrics for certain subsets of the data, perform ablation studies, compare the behaviour of two systems beyond just the performance metric, develop visualisation strategies.
Readings (if applicable)	

Session 5: 30.03.2021 **Final version**	
Aim	Discuss the key findings from the final version of the thesis
Assignment	You should prepare a final version of the thesis with all the sections completed. It should provide the discussion of the work undertaken for the project, including data collection, development of methods, experimental details, comparison with previous work, and thorough analysis. Your thesis should be modelled in the same style as an empirical research paper: either a conference proceedings (in a major conference) or a political science journal article. It should be about 15 pages long without references and appendix. Your final version of the thesis builds on your midterm milestone and should contain the following sections: 1. Abstract: should be a concise (less than 300 words) motivation of the problem, description of the aims and your contribution, and main findings. 2. Introduction: Explains the problem; why it's difficult, interesting, or important; how and why current methods fail/succeed at the problem, and explains key ideas of your approach and results. Roughly the same material as the abstract but with more space for motivation, detail, and references to existing work, and aims to capture the reader's interest. 3. Related Work 4. Proposed Method

	5. Experiments 6. Analysis 7. Conclusions: Summarise the main findings of your project and what you have learnt. Highlight your achievements, and note the primary limitations of your work. You can also describe avenues for future work. 8. Acknowledgements: list acknowledgements if any. E.g. if someone provided you with a dataset or you used someone's resources. 9. Contributions: if it's a team project you should describe the contributions of each team member. This will also be cross-referenced with the GitHub contributions. 10. References: You should use BibTeX referencing, you can use e.g. free resource like Zotero. 11. Appendix: if you wish, you can include extra details, examples, figures, results, visualisations, etc. Rule of thumb for the appendices is that you shouldn't assume that the grader to read them and hence you will not be assessed based on the content of the appendix but only on the content of the main paper. If it still makes sense to add the material to the appendix, then do it. The GitHub repository for the project should provide a fully replicable set of code and instructions, while the contribution GitHub section should reflect your progress in the course of the project. You need to send it to me by **8am, 30 March 2021**.
Readings (if applicable)	

Session 6: 13.04.2021 **Last check and project defence**	
Aim	Check any outstanding issues in the write up stage and discuss the poster concept
Assignment	You should prepare a digital poster mock up identifying the key visuals from the thesis that you intend to use. You will present and defend your project using the poster mock up. The mock up should be in your GitHub repository.
Readings (if applicable)	