

Thesis Proposals

If you like any of the topics below, *add a comment to the title* and send an email to all the reference persons stating

- your master's degree (Computer Science, Data Science, Mathematics, etc...)
- when you plan to start and finish your thesis
- how many exams you need to pass to complete your study plan and if you are allowed to start the thesis (for example some degrees have constraints on the number of credits needed to start the thesis), and the average of your grades
- if you attended the NLP course (especially if the topic of your choice is about NLP) and when you passed the exam

Attach your CV to the email.

Some of the proposals are in collaboration with other researchers (see reference section of the proposal), which implies that you might be working with him/her to a varying degree (depending on the project itself).

Proposals Currently Available

Proposal 1: Machine Learning Tools for Increasing the Quality of Machine Learning Datasets

Brief Description:

Datasets for supervised learning tasks in Natural Language Processing are often built first by collecting texts and then having human annotators deciding the labels of each text. The annotation process is prone to errors and the goal of the thesis is to devise tools to improve their quality. Examples of such tools include:

- statistics that are computed and summarised for a human supervisor, possibly automatically suggesting possible errors;
- each time a human does an annotation, it could be compared to similar annotations done in the past for the same text and alert the user in case they differ;
- a machine learning system that computes its own annotations which can be used either to help the annotator (presented after the annotator as a suggestion) or to check its work;
- a standard way to check the work of the annotators, especially when money is involved, is to let them annotate, from time to time, examples for which the correct label is known (test examples). A model that generates plausible test examples and suggest when to present them would be a useful tool.

Ideally, any tool designed should be implemented in an annotation platform (Inception: <https://inception-project.github.io/>).

skills needed: java programming

Reference:

- Elisa Sartori (elisa.sartori.2@unipd.it)
- Giovanni Da San Martino (giovanni.dasanmartino@unipd.it)

Proposal 2: Opinion Mining in Social Networks

Brief Description:

Inferring user opinions on various topics within online social networks has been a central focus of research on digital platforms. Numerous techniques have been proposed, each relying on different features and data points. While it is often straightforward to infer the opinions of highly active users or those with extreme viewpoints, accurately labeling the majority of users, particularly those who are less active or whose opinions are more moderate, remains challenging.

This project seeks to enhance both the volume and reliability of opinion inference by integrating state-of-the-art techniques with other methods such as machine learning and Bayesian inference. By applying these methods to diverse datasets, and potentially across multiple platforms, the study aims to identify the conditions under which opinion inference can be effectively extended to less active and more moderate users.

skills needed:

Reference:

- Alessandro Galeazzi for details (alessandro.galeazzi@math.unipd.it)
- Keep Giovanni Da San Martino (giovanni.dasanmartino@unipd.it) in the loop, so that I can keep my records updated.

Proposal 3: Rhetorical Analysis of Memes

Brief Description:

Memes are fun, but sometimes they are used as tools to influence and promote political agendas. Understanding the communicative mechanism behind them allows for their critical consumption. Specifically, there are mechanisms, such as using logical fallacies or leveraging of the emotions of the reader, that can be manipulatively used to potentially influence the reader. The goal of the thesis is to analyse a large dataset of memes manually annotated with such mechanisms, which we call persuasion techniques. The dataset is described here:

<https://aclanthology.org/2024.semeval-1.275.pdf>.

A thesis could be divided into two parts

1. Critically analysing all the algorithms proposed in the paper above and the ones cited in the same paper. This implies to run/reimplement all/some of the approaches.
2. Depending on how much efforts have been devoted to 1), the thesis could further develop in one of the following challenging ways (or others proposed by the student):

- the techniques (the output classes of the supervised learning task) are organised in a hierarchy. The thesis could exploit such structured information to improve the outcome of the machine learning algorithms (more info on request).
- For a significant number of memes the detection of the techniques requires combined reasoning on both on the textual and visual content of a meme. The thesis could study how to optimise standard machine learning approaches for such multimodal setting.

skills needed: good ability to install, run and fine tune LLM-based learning algorithms; good knowledge of machine learning algorithms.

Reference:

- Giovanni Da San Martino (giovanni.dasanmartino@unipd.it)

Proposal 4: Detection of Narratives in texts

Brief Description:

There is an online shared task on detection and extraction of narratives in texts (roughly speaking a narrative is a message that is explicitly or implicitly conveyed in a text). The thesis would consist in proposing a solution for such tasks (or a variation of them), which correspond to subtask 2 and 3 at

<https://propaganda.math.unipd.it/semEval2025task10/>

skills needed: good ability to install, run and fine tune LLM-based learning algorithms; good knowledge of machine learning algorithms.

Reference:

- Giovanni Da San Martino (giovanni.dasanmartino@unipd.it)

Proposal 5: Educational Learning Analytics

Brief Description:

The goal of the thesis is to analyse anonymised Moodle log data of a computer science course and make predictive models of the student outcomes from their activities. Several research questions could be tackled (only a subset of these need to be answered for a thesis): is it possible to extract indicators of the student outcomes? How relevant is each quiz/exercise? What features make a quiz/exercise a good indicator of the final performance of the student? What new exercise would improve either the prediction of the student outcomes or the outcomes itself? Can we generate ad hoc exercises that are useful for each student? Are there early signals of students' dropout? A deliverable of the thesis could be a dashboard that allows to load log data and graphically shows the result of the most interesting analysis related to any of the questions above.

Some relevant literature:

- https://drive.google.com/file/d/14bqUI7nE3jc-mclrHiX9fpAqR_w6xw7y/view?usp=sharing
- https://drive.google.com/file/d/14_u1yPiPpJRwLSDAQsSne5ztsFZ2fgJ9/view?usp=sharing

skills needed: good knowledge of machine learning algorithms, C programming and ability to build a dashboard.

Reference:

- Giovanni Da San Martino (giovanni.dasanmartino@unipd.it)

Proposal 6: Analysis of a Machine Learning Dataset

Brief Description:

The goal of the thesis is to study and measure the quality of a dataset and evaluate its impact on the accuracy of machine learning models trained on it. The dataset has millions of records. The analysis consists in developing efficient algorithms for quality evaluation and for filtering the most significant examples. As a second step, it will be analysed how different levels of quality of a dataset affect basic models, such as random forest, and more advanced ones, such as neural networks. Presence in Padova is required for this thesis.

skills needed: good knowledge of machine learning principles and algorithms

Reference:

- Martino Grasselli (grassell@math.unipd.it)
- Giovanni Da San Martino (giovanni.dasanmartino@unipd.it)

Proposal 7: Evaluating and Detecting Large Language Model-Generated vs. Human-Annotated Misinformation: Toward Improved Detection Methods

Brief Description:

Large Language Models (LLMs), such as ChatGPT, have introduced new capabilities in automated content generation. While beneficial in many contexts, these models can also be misused to produce realistic yet false information, raising serious concerns about their role in spreading misinformation. Recent studies indicate that LLM-generated misinformation may be more difficult to detect than human-written misinformation, both for automated systems and human evaluators. This increased deception potential heightens the risk to public trust and information integrity. This thesis will bridge a critical research gap by (1) providing empirical evidence on the

distinct challenges posed by LLM-generated misinformation, and (2) proposing a detection strategy tailored to the evolving nature of AI-generated deception.

Possible objectives of this thesis could be:

1. Comparative Analysis: Systematically evaluate and compare the characteristics of LLM-generated misinformation and human-annotated (written) misinformation.
2. Detection Challenges: Analyze why existing detection models struggle more with LLM-generated misinformation, identifying key linguistic features or different persuasion techniques used in both types of misinformation.
3. Method Development: Propose and implement a novel detection method or enhancement (e.g., feature-based models, fine-tuned transformers, or hybrid approaches) that improves the accuracy and robustness of identifying LLM-generated misinformation.
4. Evaluation: Rigorously evaluate the proposed detection method using benchmark datasets containing both LLM-generated and human-authored misinformation to measure its accuracy, generalizability, and effectiveness in real-world scenarios. Additionally, the creation and incorporation of newly generated corpora could further enhance the relevance of the evaluation.

Skills needed: NLP, machine learning, ability to fine-tune language models, Python programming, HuggingFace

Reference:

- Arkadiusz Modzelewski (arkadiusz.modzelewski@pja.edu.pl, to find out more about me check my website: <https://amodzelewski.com/>)
- Giovanni Da San Martino (giovanni.dasanmartino@unipd.it)

Proposal 8: Dynamics of Persuasion in Games

Brief Description:

There are several games for which the ability to influence other players is a key factor for winning: Town of Salem, Secret Hitler, Two Rooms, Among us, Diplomacy...

The goal of the thesis is to study the dynamics of persuasion in games. This requires to collect meaningful descriptions of games, i.e. logs, which can be done by:

1. finding game logs (see Diplomacy and the study made by Meta)
2. creating a simulation (or finding one available) in which agents play the game and then collecting the logs.

In both cases, the first objective of the thesis is to find successful and unsuccessful (with respect to persuasion) communication patterns and, depending on the type of logs available, how communication strategies evolve over time.

In the case a simulation is available, agents could be trained to evolve to improve their ability to win the game.

The thesis is part of a research collaboration with the CNR in Pisa, therefore it could involve collaborating with their researchers, but it is not a thesis with company internship.

Some relevant literature:

- <https://arxiv.org/abs/2504.04072>

skills needed: good ability to install, run and fine tune LLM-based learning algorithms; good knowledge of machine learning algorithms.

Reference:

- Giovanni Da San Martino (giovanni.dasanmartino@unipd.it)