FACULDADE DE
CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE D
COIMBRA

Natural Language Processing 2025/2026

Bachelor in Artificial Intelligence and Data Science

Project

Text Classification

Professors:

Isabel Carvalho
isabelc@dei.uc.pt

Patrícia Ferreira
patriciaf@dei.uc.pt

Hugo Gonçalo Oliveira
hroliv@dei.uc.pt

1 Objectives

The objective of the project for the Natural Language Processing (NLP) course is to apply the content taught in tasks related to text classification in Portuguese, using Machine Learning (ML) and *Prompt Engineering*.

The project must be developed in Python, with a justified use of libraries for NLP and ML. Throughout the semester, reports with intermediate results and conclusions must be submitted, later compiled into a final report and defended orally. The intermediate milestones will be followed by a presentation and discussion of the work with the professors.

2 Task and Data

The classification tasks aim to assign a category associated with textual data. The following datasets are suggested:

FactNews (Vargas et al., 2023) This corpus consists of sentences extracted from Brazilian Portuguese news, suitable for exploring classifiers of factuality and bias. The data were collected from three recognized news sources: Folha de São Paulo, O Globo, and Estadão. The data are available at `https://huggingface.co/datasets/franciellevargas/FactNews` or `https://github.com/franciellevargas/FactNews/tree/main`, the latter including information about the annotation process. Table 1 contains excerpts of examples from this corpus.

ID	Domain	Text	Category
465	Mundo	É difícil dizer ao certo o que está acontecendo", disse, transtornado, o estudante Jason Anthony Smith, de 19 anos.	-1 (Quote)
5499	Política	D. Luiz completou ontem 17 dias sem se alimentar.	0 (Fact)
141	Diário	Os bandidos atacaram agências bancárias, bases policiais e prédios públicos com bombas e tiros.	1 (Bias)

Tabela 1: Excerpts of examples from the FactNews corpus.

HateBR (Vargas et al., 2022) This corpus consists of comments extracted from posts by Brazilian politicians on *Instagram*, suitable for exploring classifiers of offensive language. The data were manually annotated by three experts, whose annotations are made available together with the dataset at `https://huggingface.co/`

`datasets/franciellevargas/HateBR` or `https://github.com/franciellevargas/HateBR/tree/main`. Table 2 contains excerpts of examples from this corpus.

ID	Text	Category
4839	Eita homem lindo! Por mim era nosso presidente!	0 (Not Offensive)
293	Que faz este cara ainda morando fora da prisão?	1 (Offensive)

Tabela 2: Excerpts of examples from the HateBR corpus.

FakeWhatsApp.Br (Cabral et al., 2021) This corpus consists of anonymized messages extracted from public political campaign groups on *WhatsApp*, suitable for exploring classifiers of misinformation. The messages considered viral were manually annotated. The data are available at `https://github.com/cabrau/FakeWhatsApp.Br`, and the file to be used is *data/2018/fakeWhatsApp.BR_2018.csv* (the category -1 represents unannotated data). Table 3 contains excerpts of examples from this corpus.

ID	Text	Category
5921081908262978616	Bom dia E viva o Brasil Brasil acima de todos e Deus acima de tudo	0 (No Misinformation)
8371047632500103882	Esse cara deu um soco no baço e um puxão/beliscão para agravar o ferimento. Compartilhe até que seja identificado.	1 (Misinformation)

Tabela 3: Excerpts of examples from the FakeWhatsApp.Br corpus (*data/2018/fakeWhatsApp.BR_2018.csv*).

The availability of information in the mentioned repositories, e.g., models, code, features, and pre-processed text, does not exempt compliance with the academic fraud regulations.

For more information about the evaluation, datasets, and common approaches for these tasks, it is suggested to consult the references at the end of this document, including the course bibliography (Alammar and Grootendorst, 2024, Caseli and Nunes, 2024, Einsenstein, 2018, Jurafsky and Martin, 2025, Tunstall et al., 2022).

3 Milestones

The project is divided into three main milestones:

Milestone 1: Development of a rule-based text classification system, which will include:

- Data analysis and exploration of linguistic knowledge extracted using different libraries, e.g., tokens, n-grams, grammatical functions, entities, sentiment, relations, etc.;
- The definition and justification of rules that leverage this knowledge for text analysis and respective categorization;
- Performance analysis on the mentioned dataset to draw conclusions about the most and least useful knowledge for the task.

Milestone 2: Development of a text classification system using ML methods, which will include:

- Training one or more ML models based on text content and/or features derived from linguistic knowledge extracted using different libraries, e.g., grammatical functions, entities, sentiment, relations, semantic similarity, etc.;
- Performance analysis on the mentioned dataset to draw conclusions about the best models and the most useful features.

Milestone 3: Development of a text classification system using *Prompt Engineering*, which will include:

- The exploration of *Prompt Engineering* in at least one open language model based on transformers, with a maximum of 8B parameters. The models must be properly documented;
- Performance analysis on the mentioned dataset to draw conclusions about the best models, comparison with the results of the previous milestones, error analysis, and description and evolution of the prompts used.

Appropriate evaluation metrics for the problem must be defined and properly justified.

4 Submissions

The development and results of the first two milestones will be presented in mini-reports, with a maximum of 3 pages each, excluding references. These reports must describe the data processing workflow, the experimentation and evaluation process, including preliminary conclusions and justification of the best approach and choice

of metrics. Along with the mini-reports, all files used in their preparation must be submitted, including code. In the class following each intermediate submission, there will be a mandatory presentation and discussion with the professor.

The work will culminate in the final milestone, which must include all developed code files and a final report, with a maximum of 8 pages, excluding references, compiling the results of the three milestones in an organized manner. The following structure is suggested:

- Introduction;
- Detailed description of the experiments carried out, including features and approaches explored, their performance, and error analysis;
- Summary, including main challenges and conclusions;
- Bibliographic references.

The final submission will also be subject to an oral presentation, scheduled during the last week of classes of the semester. Information on dates and submissions is presented in Table 4.

The files associated with each milestone must be submitted through the course page on Inforestudante. The developed/used models themselves must not be submitted.

5 Evaluation Criteria

The NLP project accounts for 50% of the final grade of the course, i.e., 10 points out of 20, distributed as follows:

- Milestone 1: 2 points, 10% of the final grade of the course;
- Milestone 2: 2 points, 10% of the final grade of the course;
- Milestone 3: 6 points, 30% of the final grade of the course.

The evaluation will take into consideration the following main aspects:

- Fulfillment of requirements and demonstrated learning;
- Relevance, scope, originality, performance, and complexity of the experiments carried out;

Attention: Plagiarism will not be tolerated!

Milestone	To be delivered	Date
Group and corpus definition	Fill in Google Sheets (`https://tinyurl.com/grupos-pln`)	September 26, 2025
Milestone 1	• Mini-report (max. 3 pages) • Code and files used	October 17, 2025
Milestone 1 Presentation	5 minutes (presentation) + 5 minutes (discussion)	Week of October 20, 2025
Milestone 2	• Mini-report (max. 3 pages) • Code and files used	November 14, 2025
Milestone 2 Presentation	5 minutes (presentation) + 5 minutes (discussion)	Week of November 17, 2025
Milestone 3	• Final report (max. 8 pages) • Code and files used	December 12, 2025
Defense	5 minutes (presentation) + 10 minutes (discussion)	Week of December 15, 2025

Tabela 4: Timeline of submissions and presentations

- Quality of the conclusions and their discussion;
- Presentation and oral discussion of the work.

Additional notes:

- The work must be exclusively authored by the group, with each member assuming responsibility for the materials presented. The use of writing support tools or code reuse must be explicitly stated and carried out with a critical approach, including a review of the content and clear mention in the reports of their usage and the prompts executed, ensuring full transparency in the process;
- It is mandatory to cite all sources used, including books, articles, web pages, language models (LLMs), or any other format through which external work is used;
- Attendance, presentation, and oral discussion during the presentations of Milestones 1 and 2, as well as the final defense, must involve all group members,

with each one being responsible for answering questions related to the work developed;

- Although the project is carried out in groups, the final grade may differ among group members, depending on their individual performance.

References

Alammar, J. and Grootendorst, M. (2024). *Hands-on large language models: language understanding and generation*. "O'Reilly Media, Inc.".

Cabral, L., Monteiro, J. M., da Silva, J. W. F., Mattos, C. L. C., and Mourao, P. J. C. (2021). FakeWhastApp.BR: NLP and Machine Learning Techniques for Misinformation Detection in Brazilian Portuguese WhatsApp Messages. In *Proceedings of the 23rd International Conference on Enterprise Information Systems, ICEIS. 2021*, pages 63–74.

Caseli, H. M. and Nunes, M. G. V., editors (2024). *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*. BPLN, 2 edition.

Einsenstein, J. (2018). *Natural Language Processing*. MIT press.

Jurafsky, D. and Martin, J. H. (2025). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, with Language Models*. 3rd edition. Online manuscript released August 24, 2025.

Tunstall, L., Von Werra, L., and Wolf, T. (2022). *Natural language processing with transformers*. "O'Reilly Media, Inc.".

Vargas, F., Carvalho, I., Rodrigues de Góes, F., Pardo, T., and Benevenuto, F. (2022). HateBR: A large expert annotated corpus of Brazilian Instagram comments for offensive language and hate speech detection. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, pages 7174–7183, Marseille, France. European Language Resources Association.

Vargas, F., Jaidka, K., Pardo, T., and Benevenuto, F. (2023). Predicting sentence-level factuality of news and bias of media outlets. In Mitkov, R. and Angelova, G., editors, *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 1197–1206, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.