A Hitchhiker's Guide to NLP Semester Project

Mehmet Can Yavuz, PhD

This document provides a detailed, step-by-step framework for the successful execution of your semester-long technical project. This project is designed to be a capstone experience, simulating the end-to-end lifecycle of a data science or machine learning initiative, from initial problem formulation to final dissemination of results. Adherence to this structured guideline is crucial for meeting the project milestones and achieving a high standard of academic and technical rigor. The project is divided into three primary phases, each with specific deliverables and grading criteria designed to build upon the previous stage.

Info

Why This Matters. This project is not just another homework assignment: it is designed as your first opportunity to simulate the full research and development cycle. You will practice identifying a real-world problem, finding suitable data, implementing models, and then explaining your findings. These are the same steps followed in both academic research and industry projects. Treat this guideline as a roadmap to completing your first "mini-thesis."

Project Policies and Resources

Collaboration Policy and Academic Integrity This is an individual project. The core intellectual work—including model implementation, experimental design, analysis, and writing—must be your own.

However, academic research is often collaborative. Therefore, limited collaboration is permitted for specific, non-core components of the project. Permissible areas for collaboration include:

- **Brainstorming:** Discussing ideas, potential datasets, or high-level approaches with classmates.
- **Literature Search:** Sharing interesting papers or resources you discover.

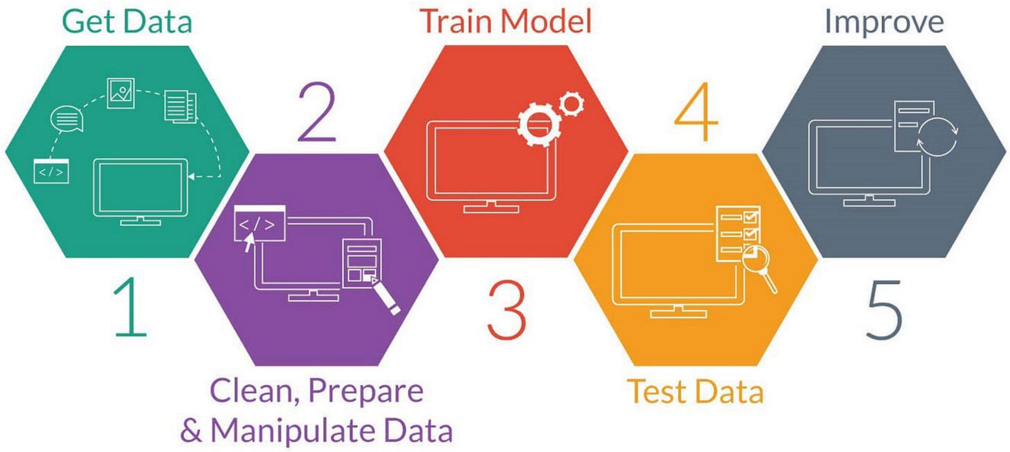

Figure 1: The typical workflow for a machine learning project, starting from data acquisition to model improvement.

Important

Collaboration Statement. You must explicitly declare all collaborations. In your final report, include a "Collaboration Statement" section detailing who you collaborated with and the specific nature of the collaboration. Failure to disclose collaboration will be considered a violation of academic integrity.

Part I: Project Proposal (Weight: 5%)

The initial phase of your project lays the intellectual and strategic foundation for all subsequent work. Success in this stage is not merely about selecting a topic but about defining a clear, compelling, and methodologically sound research direction. This phase culminates in the submission of a formal project proposal, which will serve as the definitive blueprint for your investigation. A well-crafted proposal demonstrates foresight, a thorough understanding of the problem domain, and a viable plan for execution, thereby setting the trajectory for the entire semester.

Section 1: Defining Your Research Trajectory: Topic and Dataset Selection

1.1 The Primacy of the Problem Your project must begin not with a dataset, but with a compelling problem or a well-articulated question. The primary objective is to define the problem you are trying to solve in a real-world context. The first step is to select a domain that genuinely interests you—within the course content. Intrinsic motivation is a critical factor for sustaining the high level of effort required for a semester-long project.

Info

Common Pitfall. Many students start by choosing a dataset they find online and then try to force a research question to fit it. Instead, flip the order: begin with a problem you care about (e.g., misinformation on social media, hate speech detection, or translation). The dataset is the tool, not the driver.

1.2 The Search for Data With a well-defined problem in hand, the search for a suitable dataset can commence. Utilize established, publicly available datasets from reputable platforms such as Kaggle, the UCI Machine Learning Repository, Google's Dataset Search, or data journalism sites like FiveThirtyEight. These datasets are often pre-cleaned, well-documented, and structured for machine learning tasks. This allows you to concentrate your efforts more on the modeling and analysis phases of the project.

Info

The relationship between a project's problem statement and its dataset is symbiotic. While a compelling problem should ideally guide the search for data, the practical constraints and characteristics of available data will often necessitate a refinement of the initial problem statement.

Section 2: The Literature Review

Purpose and Significance A literature review is a mandatory, foundational component of your proposal. It is a critical, analytical survey of existing scholarly work relevant to your topic. Its objectives are to demonstrate topical familiarity, develop a theoretical and methodological framework, and position your work within an ongoing academic conversation by identifying a concrete gap your project will address.

A Five-Step Process for a Rigorous Review A systematic approach ensures quality:

1. **Search for Relevant Literature:** Use carefully chosen keywords and academic databases (e.g., Google Scholar, ACM DL, IEEE Xplore). Prioritize peer-reviewed sources.
2. **Evaluate and Select Sources:** Judge relevance and credibility by asking what problem is addressed, which methods are used, key results, and strengths/weaknesses.
3. **Identify Themes, Debates, and Gaps:** Synthesize across sources; highlight trends, conflicts, and unanswered questions to locate your gap.
4. **Outline the Structure:** Organize thematically, methodologically, or theoretically to serve your argument rather than listing works chronologically.
5. **Write the Review:** Synthesize sources within paragraphs, showing consensus and debates, and argue why your project is the logical next step.

Topic	Authors & Year	Dataset (with Link)
Author Profiling	Koppel et al. (2009) Rangel et al. (2017)	Blog Authorship Corpus PAN Author Profiling Dataset
Emotion Detection	Mohammad & Turney (2013) Demszky et al. (2020)	NRC Emotion Lexicon GoEmotions
Fake News Detection	Pérez-Rosas et al. (2018) Wang (2017)	FakeNewsNet LIAR Dataset
Hate Speech Detection	Davidson et al. (2017) Fortuna & Nunes (2018)	Hate Speech and Offensive Language Dataset HASOC Shared Task Dataset
Machine Translation	Bahdanau et al. (2014) Vaswani et al. (2017)	IWSLT German–English WMT Shared Task Datasets
Named Entity Recognition	Tjong Kim Sang & De Meulder (2003) Chiu & Nichols (2016)	CoNLL-2003 Dataset OntoNotes 5.0
News Categorization	Blei et al. (2003) Chang et al. (2009)	20 Newsgroups AG News Corpus
POS Tagging	Marcus et al. (1993) Plank et al. (2016)	Penn Treebank Universal Dependencies Treebanks
Question Answering	Rajpurkar et al. (2016) Kwiatkowski et al. (2019)	SQuAD v1.1/v2.0 Natural Questions (Google)
Sentiment Analysis	Maas et al. (2011) Socher et al. (2013)	IMDb Movie Reviews Stanford Sentiment Treebank (SST)
Spam Classification	Androutsopoulos et al. (2000) Metsis et al. (2006)	LingSpam Dataset Enron Email Dataset
Word Sense Disambiguation	Navigli (2009) Raganato et al. (2017)	SemCor SemEval WSD Benchmarks

Info

Dataset Access Note. Some resources (e.g., Penn Treebank, OntoNotes 5.0, CoNLL-2003 English) may require LDC or subscription access or maybe broken links. If access is unavailable, use open alternatives such as *WikiANN (Pan-X)* for NER, *Universal Dependencies* for POS/Parsing, or *WNUT-2017* for emerging/rare entities.

A thorough review yields a technical blueprint: it identifies state-of-the-art methods to implement, metrics to target, and historically grounded baselines to beat (e.g., CNN/ResNet for image classification; SVM with hand-engineered features as a baseline).

Info

Tip. Think of your literature review like joining a dinner conversation. Before you speak, you must know what has already been said, who disagrees with whom, and what questions are still unresolved. Your project should position itself as one more "voice" in that conversation.

Section 3: Crafting a Project Proposal

An effective proposal must clearly articulate the essential elements of the research endeavor by answering three fundamental questions in a coherent narrative. It should specify **what** the problem is and the ultimate goal of the project, **why** the problem matters in terms of scholarly or practical significance, and **how** the research will be conducted through a well-defined technical approach and methodology. In this sense, the proposal should be treated not merely as a plan but as a persuasive contract that establishes the scope, expectations, and direction of the project.

Once your proposal is approved, you move from "planning mode" into "engineering mode." This is where ideas are tested against data. The quality of your preprocessing and preparation will directly shape the performance of your models—think of it as cleaning and sharpening your tools before starting the actual construction.

Phase	Key Task	Deliverable	Target
Proposal (5%)	Define topic & dataset	1-paragraph description of problem and data	Week 2
	Literature review	Annotated bibliography (5+ sources)	Week 2–3
	Deadline	Research Question Video (max 3 mins) + Research Highlights Sheet	Week 3
Modeling (15%)	Data preprocessing	Cleaned dataset + EDA Analyses	Week 4–6
	Implementation	Model Implementation	Week 6–9
	Hyperparameter Opt.	short HPO write-up	Week 9–10
	Evaluation	Metrics summary (baseline vs. best HPO)	Week 11
	Deadline	Research Evaluation Video (max 4 mins) + Research Highlights Sheet	Week 11
Synthesis (25%)	Interpretation of results	Error analysis (one-to-one project meeting)	Week 12–13
	Final submission	Final Report (max 6 pages) + Presentation Video (max 5 mins) + GitHub repo	Week 14
	Project presentation	10–15 min in-class presentation with slides	Week 14

Info

How to Read This Table. Each row represents both a milestone (what you should have done by a certain week) and a grading checkpoint (how your effort is evaluated). Do not treat these deadlines as last-minute submissions; instead, see them as markers to keep your semester-long project on track. Missing one step will make the next phase much harder.

Part II: Modeling (Weight: 15%)

The modeling phase represents the technical core of the project, where students transform raw data into actionable insights through preprocessing, modeling, and systematic evaluation. Deliverables here are not merely about code or metrics, but about demonstrating an end-to-end pipeline with reproducible results.

Section 4: Data Preparation and Preprocessing

4.1 A Systematic Approach Data is the foundation of every model. Students must provide:

- Cleaned and well-documented datasets.
- Exploratory Data Analysis (EDA) visualizations and summary statistics.
- Justification for preprocessing choices (e.g., normalization, handling missing data).

Info

The quality of preprocessing directly influences model performance. Reproducibility requires that every step of cleaning and transformation be fully documented.

Exploratory Data Analysis (EDA): Uncovering Initial Insights

Info

Core EDA Techniques EDA involves univariate and multivariate analysis.

Univariate (one variable at a time).

- **Numerical:** Descriptive statistics (mean, median, mode; variance, SD, IQR); histograms, density plots, box plots.
- **Categorical:** Frequency counts; bar charts.

Bivariate/Multivariate (relationships).

- **Correlation Analysis:** Pearson's r for numerical pairs; correlation matrix (heatmap).
- **Visualizations:** Scatter plots for continuous pairs; grouped box plots for continuous vs. categorical; cross-tabulations for categorical pairs.

Section 5: Model Implementation

5.1 Baseline Models Each student must:

- Establish a reproducible baseline.
- Store all code in a structured GitHub repository with README, `requirements.txt`, and usage examples.

Section 5A: Hyperparameter Optimization (HPO)

Objective Systematically search the configuration space to improve model performance under a fixed compute budget while preserving reproducibility and fair evaluation.

Definition

Hyperparameter Optimization (HPO). The process of selecting hyperparameters (learning rate, regularization strength, architecture sizes, etc.) that are not learned from data directly but critically affect model performance.

Section 6: Evaluation and Reporting

6.1 Metrics and Interpretability Deliverables include:

- A comprehensive evaluation with appropriate metrics (accuracy, F1, BLEU, etc.).
- An error analysis identifying common failure cases.
- Clear documentation of hyperparameters and training process.

At the end of this phase, students will submit:

- **Video Presentation (max. 4 mins):** Demonstrating preprocessing, model design, and evaluation results.
- **Research Highlights Sheet:** Summarizing the dataset pipeline, chosen models, baseline results, and key insights.

Info

Why Evaluation Matters. Reporting only accuracy numbers is like saying "my car works" without explaining how fast it goes, how safe it is, or how fuel-efficient it might be. A strong evaluation shows the full picture: what your model is good at, where it fails, and how those results connect back to the original research question.

Guidelines for Video Presentations and Research Highlights Sheet

Video Presentation (max. 4 minutes)

Your video presentation is a short, structured, and professional recording where you walk the audience through your project progress. It is not a casual recording but a mini-research talk.

General Requirements

- **Duration:** Maximum 4 minutes (strict).
- **Format:** Recorded screen presentation (slides + voiceover; webcam optional).
- **Clarity:** Speak slowly and clearly; avoid technical jargon unless explained.
- **Visuals:** Slides should be readable (large fonts, minimal text, supportive visuals).

Recommended Structure

1. **Introduction (30–45 seconds):** Introduce yourself and your project title. Clearly state the problem or question your project addresses. Briefly explain why this problem is important.

2. **Dataset & Preprocessing (45–60 seconds):** Describe the dataset(s) used (size, source, characteristics). Mention preprocessing steps (cleaning, normalization, handling missing values, etc.). Show one visual example (e.g., distribution plot, word cloud).

3. **Modeling Approach (60–75 seconds):** State your baseline model (e.g., Naïve Bayes, Logistic Regression) if applicable. Justify why these models are appropriate. Include 1–2 clear diagrams or charts (architecture figure, pipeline diagram).

4. **Evaluation & Results (45–60 seconds):** Present your evaluation metrics (accuracy, F1, BLEU, etc.). Baseline model. Show at least one table or plot highlighting performance. Mention any surprising results or challenges.

5. **Conclusion (30–45 seconds):** Summarize the main findings (what worked, what didn't). Briefly mention next steps (error analysis, future improvements). End with a clear takeaway message.

Research Highlights Sheet

The highlights sheet condenses your work into a visually appealing, easy-to-read research flyer. Think of it as an executive summary for someone who doesn't have time to read your full report.

General Requirements

- **Length:** Maximum 1 page (single column or two-column layout).
- **Format:** PDF, professional academic style.
- **Visuals:** Use at least two visuals (diagram, chart, or table).

Structure

Definition

1. **Title and Authors:** Project title, your name, course name.
2. **Problem Statement (2–3 sentences):** Clearly articulate the problem or question. Highlight why it matters (academic or practical relevance).
3. **Dataset (2–3 sentences + table/visual):** Describe dataset(s) used (source, size, key properties). Include a small descriptive table or figure (e.g., class distribution).
4. **Methods (short paragraph or bullet points):** Baseline model(s). Key preprocessing/feature engineering steps. Use a pipeline diagram if possible.
5. **Results (figure + 2–3 sentences):** Key metrics (baseline). Present in a concise table or bar chart.
6. **Key Insights (3–4 bullet points):** Lessons learned, surprising findings, limitations.
7. **Future Work (1–2 sentences):** What you plan to do next.
8. **References (optional, 2–3 max):** Cite any major datasets/papers you directly used.

Design Tips

- Use visuals to reduce text: charts > paragraphs.
- Ensure readability at a glance: key numbers and insights should stand out.
- Maintain professional formatting: consistent fonts, aligned sections.

Info

Every milestone (proposal, modeling, synthesis) requires both a concise video presentation and a one-page Research Highlights Sheet. This dual format develops both communication and scientific writing skills.

Part III: Synthesis and Dissemination (Weight: 25%)

The synthesis phase ensures that technical work is translated into scientific contributions. Students move from raw results toward interpretation, presentation, and dissemination.

Section 7: Interpretation of Results

7.1 Beyond Metrics Students must demonstrate:

- Deep error analysis (both quantitative and qualitative), including identification of common misclassifications or systematic weaknesses.

- Insights into the broader implications of their findings for the research problem and potential applications.
- Critical reflections on methodological limitations and proposed avenues for improvement or extension.

Info

Synthesis transforms "numbers" into "knowledge." Successful projects do not stop at reporting accuracy values; they explain *why* results matter and what they reveal about the underlying problem.

7.2 One-to-One Feedback Session At this stage, each student (or project group) will schedule a **one-to-one meeting with the professor**.

- The meeting will focus on discussing the interpretation of results, identifying strengths and weaknesses in the analysis, and refining the narrative for the final report.
- Students are expected to come prepared with: (i) a draft error analysis, (ii) evaluation plots/tables, and (iii) specific questions or concerns.
- Feedback provided in this session must be directly incorporated into the final synthesis deliverables.

Info

Numbers vs. Knowledge. Accuracy, F1, BLEU, or any metric is just a number. Your task is to turn that number into knowledge: Why did the model succeed here? Why did it fail there? What does that tell us about the underlying dataset or problem? This is the true intellectual contribution of your project.

Final Report and Analysis Guidelines

Final Report Structure (Maximum 6 pages)

Your final report should be structured like a formal academic conference paper. Adherence to this structure is key for clarity and professionalism. Using a standard LaTeX template, such as those for ACL or IEEE conferences, is highly recommended.

1. **Title and Author:** Your project title and name.
2. **Abstract (1 paragraph):** A concise summary stating the problem, methods, main results, and key conclusion.
3. **Introduction (~1 page):** Motivate the problem, state your research question(s), briefly describe your approach, and summarize your key findings.
4. **Related Work (~0.5-1 page):** Situate your project within the existing literature and clearly state how your work differs or builds upon prior research.

5. **Methodology (~1.5 pages):** Describe your dataset, preprocessing steps, model architectures (with equations or diagrams), and experimental setup (training procedure, hyperparameters, metrics).

6. **Results and Analysis (~1.5 pages):** Present your main results in clear tables and figures. This section must include your deep error analysis.

7. **Discussion and Future Work (~0.5 page):** Interpret your results, acknowledge the limitations of your work, and suggest concrete directions for future research.

8. **Conclusion (1 paragraph):** Restate your main findings and their significance.

9. **References:** List all cited works in a consistent format.

Actionable Techniques for Deep Error Analysis

A strong project moves beyond reporting metrics. Error analysis is where you demonstrate a deep understanding of your model's behavior. Here are techniques you should use:

Quantitative Analysis

- **Confusion Matrix:** Visualize which classes your model confuses most often to reveal systematic misunderstandings.
- **Per-Class Metrics:** Report precision, recall, and F1-score for each class. This is crucial for imbalanced datasets.
- **Performance on Data Slices:** Analyze your model's performance on different subsets of your data (e.g., short vs. long sentences; texts with vs. without negation; examples with rare words).

Qualitative Analysis

- **Manual Inspection of Errors:** Sample 20–30 misclassified examples and categorize the source of error (e.g., sarcasm, ambiguity, domain-specific language, data labeling error).
- **Case Studies:** Select 2–3 particularly interesting examples (correct, incorrect, surprising) and analyze them in-depth in your report.

Section 8: Final Deliverables

The final stage of your project is about *communication.* You already know the results, but the real skill lies in explaining them clearly to others—whether classmates, professors, or a broader audience. Think of this as writing the conclusion to a research paper and delivering a short conference talk rolled into one.

- **Final Report:** Maximum 6 pages, written in a professional academic style (with figures, tables, references).
- **GitHub Repository:** Public, fully documented code, dataset references, and reproducibility notes.
- **Presentation Video (max. 5 minutes):** A polished video presenting problem, methods, results, and conclusions.
- **Research Highlights Sheet:** Executive-style summary that condenses contributions for a broad audience.
- **In-Class Presentation:** 10–15 minutes with slides, followed by Q&A.

Table 1: Phase I: Project Proposal (5%)

Criteria	Excellent (top points)	Good (mid points)	Satisfactory (min pass)	Needs Improvemen (low points)
Problem Definition (1 pts)	Research question is clear, focused, and compelling. Motivation is well-articulated.	Research question is clear but could be more focused.	Problem is stated but is vague or overly broad.	No clear research question.
Dataset Selection (1 pts)	Dataset is highly appropriate for the problem, well-described, and properly sourced.	Dataset is appropriate, but description lacks some details.	Dataset is chosen, but its relevance is unclear.	No dataset identified or it is unsuitable.
Literature Review (2 pts)	Deep engagement with 5+ high-quality sources; clearly identifies a research gap.	Engages with 3–4 sources; gap is implied but not explicit.	Lists sources with minimal synthesis or analysis.	Missing or irrelevant literature review.
Communication (1 pts)	Video and highlights sheet are professional, concise, and clear.	Communication is clear but could be more polished.	Key information is present but poorly organized.	Presentation is unclear and unprofessional.

Table 2: Phase II: Modeling (15%)

Criteria	Excellent (top points)	Good (mid points)	Satisfactory (min pass)	Needs Improvemen (low points)
Data Preprocessing & EDA (4 pts)	Thorough, well-justified preprocessing; EDA yields clear insights; fully reproducible pipeline.	Adequate preprocessing; EDA present but surface-level.	Basic cleaning without justification; little/no EDA.	Data used as-is with evident issues.
Model Implementation & Hyper-parameter Optimization(6 pts)	Baseline model correctly implemented and justified; well-structured repo.	Models implemented, but justification or code structure is weak.	Only a baseline is implemented or code is messy/undocumented.	Code is non-functional or missing.
Evaluation (3 pts)	Appropriate metrics; rigorous model comparison; results clearly presented.	Appropriate metrics but comparison is basic.	Metrics are incorrect/inappropriate.	No evaluation or results presented.
Communication (2 pts)	Video and highlights sheet clearly explain pipeline, models, and results with insightful visuals.	Clear presentation but lacks depth or strong visuals.	Hard to follow or missing key components.	Unclear, incomplete, or missing deliverables.

Table 3: Phase III: Synthesis & Final Deliverables (25%)

Criteria	**Excellent** (top points)	**Good** (mid points)	**Satisfactory** (min pass)	**Needs Improvemen** (low points)
Analysis (3 pts)	Exceptional error analysis, insightful discussion, and critical reflection on limitations.	Results analyzed, but discussion is surface-level.	Metrics reported with minimal interpretation or error analysis.	Little to no analysis.
Clarity & Structure (2 pts)	Flawless academic structure; clear, concise, professional writing; excellent figures/tables.	Well-structured and clear with minor issues.	Confusing structure; writing unclear in places.	Poorly written and organized.
GitHub Repo (5 pts)	Clean, well-documented, fully reproducible (single script/command); excellent README.	Code runs but requires manual steps; good documentation.	Runs but messy/undocumented/hard to set up.	Missing, non-running, or not reproducible.
Final Presentation (15 pts)	Engaging, professional, and clear; mastery of topic; excellent Q&A handling.	Clear, well-practiced; adequate Q&A.	Disorganized or hard to follow; struggles with Q&A.	Unprepared, unclear, or unprofessional.