From Chaos to Clarity

Evaluating Methods for Classifying Unstructured Ticket Data

Thibault Giesbertz - BSc AI Thesis

Introduction & Methods

Introduction

- Helpdesk tickets are often noisy, inconsistent, and lack structured labels.
- This unstructured format hinders automation: it's hard to retrieve solutions, assign priorities, or route tickets effectively.
- Manual labeling is time-consuming and limited in scale.
- This project explores how to classify tickets under low-resource conditions, using only a small set of labeled examples.
- We compare supervised models and clustering approaches to assess classification effectiveness, not only in terms of accuracy (macro-averaged F1), but also in practical aspects such as implementation complexity, scalability, and interpretability in real-world enterprise environments.

Examples of Real Helpdesk Tickets

Below are artificially replicated examples based on real tickets, illustrating the ambiguity of user-submitted issues and the noise introduced by system-generated or logistical messages.

Title	Description	Issue Type
Laptop X	—	*Too brief / Underspecified*
X has no mail on phone	—	*Ambiguous intent*
File server from the AVD	—	*Incomplete and noisy*
Error message when logging into AVD	"X gets an error message when trying to log in"	*Vague error / Lacks details*
Password reset	"User X needs their password reset."	*Clear and actionable (ideal)*

Research Question

"What is the most effective method for classifying unstructured helpdesk tickets in low-resource settings: supervised classification or unsupervised clustering?"

Effectiveness is evaluated by

- Macro F1-score
- Scalability, interpretability, and ease of deployment

Dataset

- Real-world data from Dutch Technology eXperts (DTX)
- 484 (*554 with Augmented set*) labeled samples, 7 categories
- Three variants:
  - **Cleaned** where private information was removed
  - **Stemmed** & stop words removed (w & w/o negation)
  - **Augmented** via back-translation where we artificially generate new tickets by translating existing tickets from Dutch -> English, English -> Dutch

Class Distribution in dataset
"Back-translation improved balance by synthetically increasing underrepresented categories."

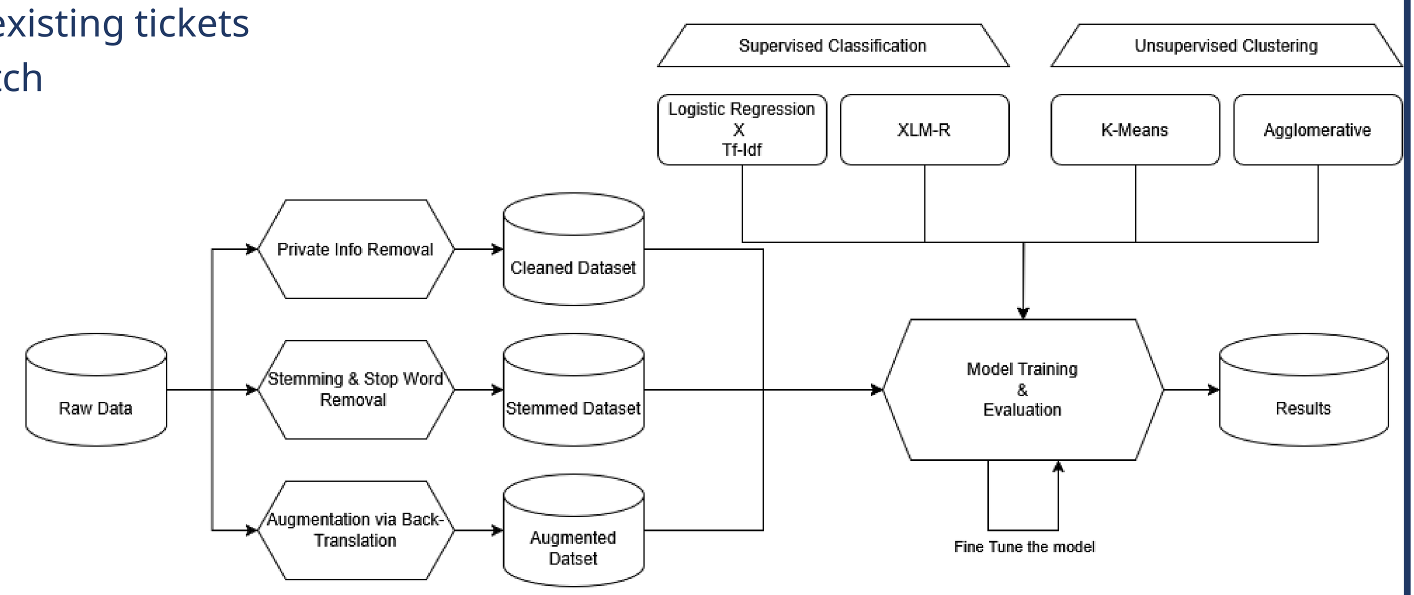
Model & Embeddings

Supervised

- Logistic Regression (TF-IDF)
- XLM-R (Transformer, multilingual)

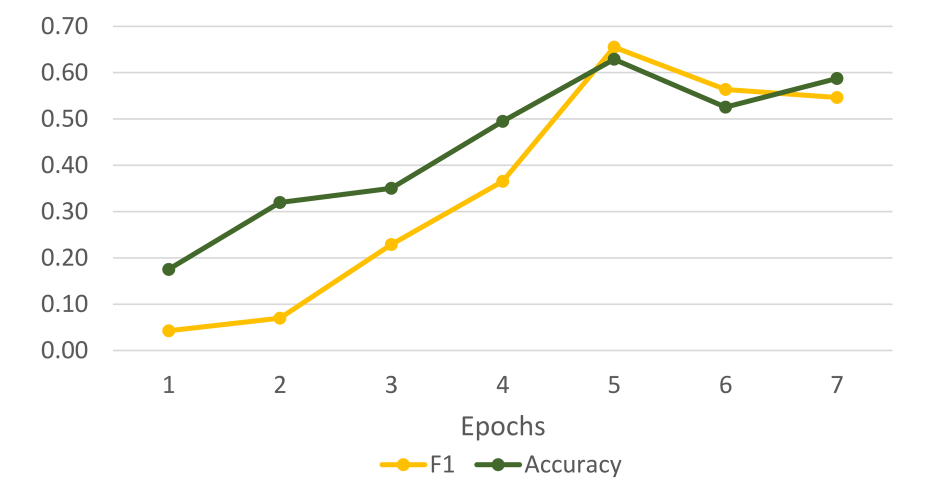
Unsupervised

- K-means & Agglomerative clustering
- Sentence-transformer embeddings (MiniLM-L6-v2) & TF-IDF
  - Both run with fixed k = 7 to match label structure



Process Schema
"Overview of the experimental pipeline: from labeled data through model training, evaluation, and iteration."

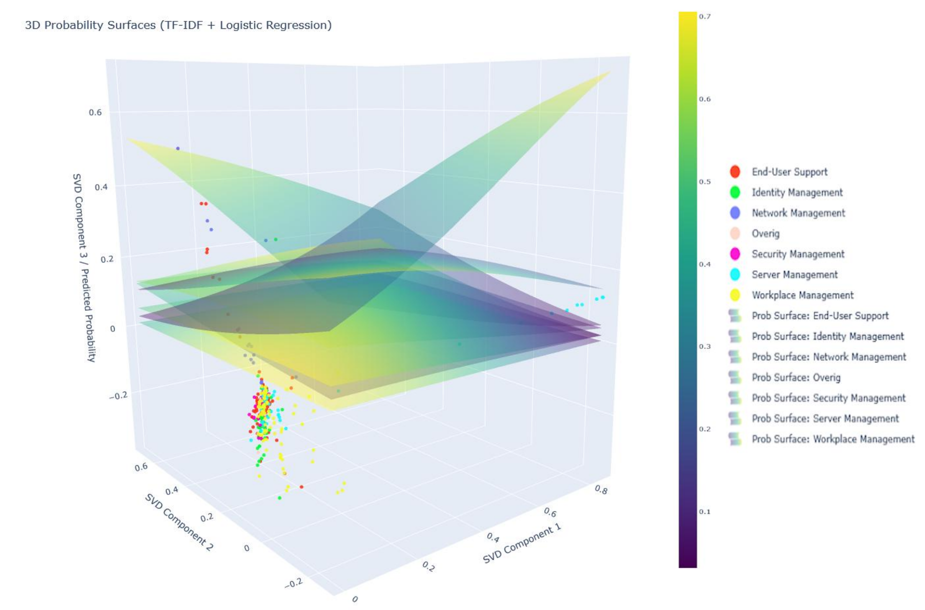
Results



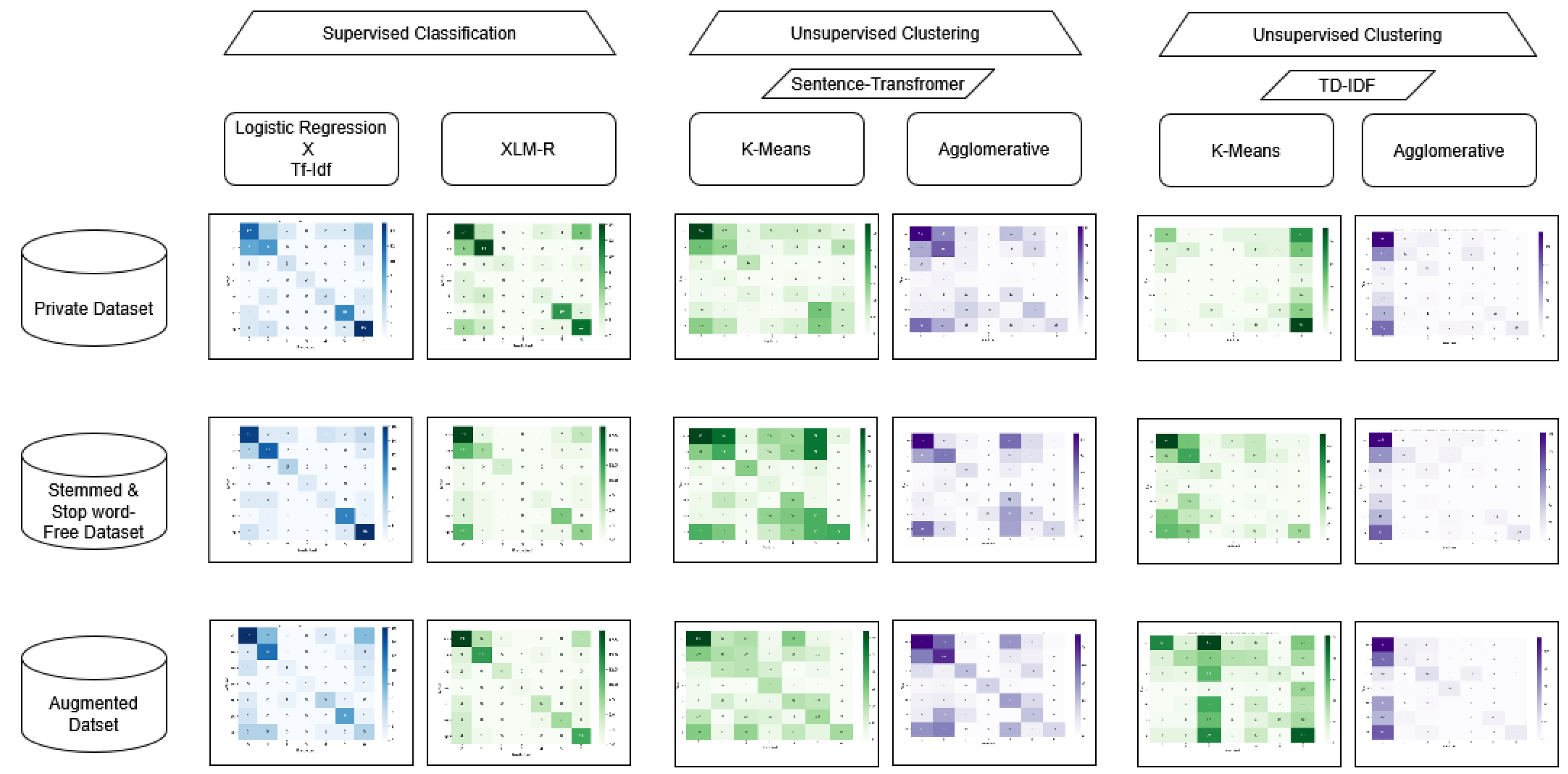
XLM-R Training Curve
"XLM-R gradually improves across epochs, indicating benefit from extended training on limited data."

Dataset	Logistic Regression	XLM-R	K-Means	Agglomerative
Cleaned	0.62	0.56	0.31	0.22
Stemmed	0.67	0.53	0.31	0.27
Stemmed (w/o negation)	**0.70**	0.48	0.32	0.32
Augmented	0.51	**0.66**	**0.34**	**0.42**

F1 Score Summary Table
"Macro F1-scores across models and datasets show supervised methods outperform unsupervised clustering."



3D Logistic Regression Plane Visualization
"Visualizing how the linear model separates classes using top features in reduced dimensions."



Confusion Matrices
"Logistic Regression demonstrates strong precision and recall on most classes, even with limited labeled data."

Interpreting the Results

- **F1-score** gives a standardized summary of model accuracy across all categories.
- **Confusion matrices** reveal detailed behavior:
  - Diagonal dominance shows successful classification.
  - Clustered errors highlight ambiguity between categories.
  - Flat rows or columns signal overfitting or category collapse.
- **Clustering models** reveal internal structure, but this structure often diverges from the intended label schema.

Discussion & Future Work

Key Insights

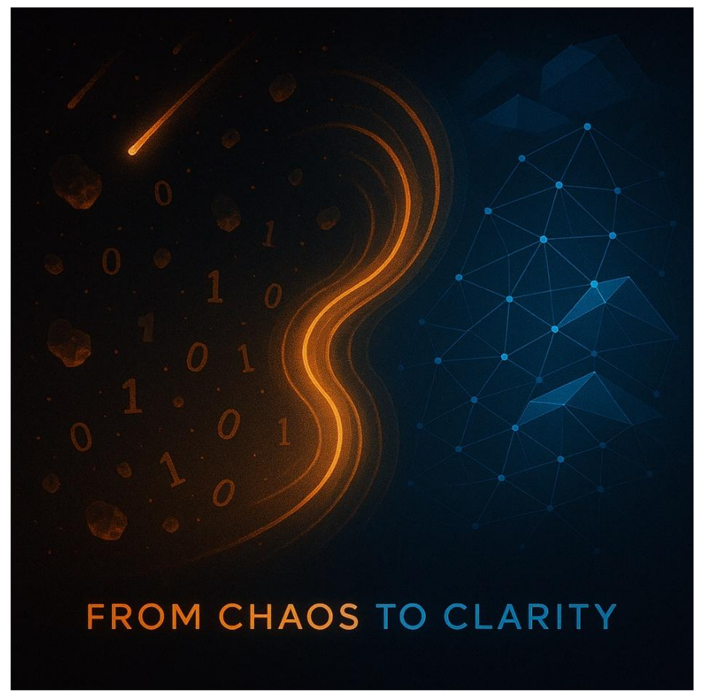
- **Supervised models (esp. Logistic Regression) outperform clustering, even with small datasets**
- **Clustering produced meaningful patterns but lacked alignment with practical categories**
- **Sentence embeddings improved clustering, but still fell short of supervised performance**

Limitations

- Small labeled set: 484 samples
- Ambiguous class boundaries (e.g. Workplace vs End-User)
- Clustering constrained to 7 fixed categories
- XLM-R not compared to larger LLMs (e.g., GPT-4)

Future Work

- Pseudo-labeling with Logistic Regression already underway (→ ~1500 labels)
- Shift to XLM-R once label quality stabilizes
- Explore guided clustering (keyword-based)
- Evaluate open-source LLMs when infrastructure allows



**Supplementary Material such as Full metrics, calibration curves, word clouds, and training logs available on GitHub (QR)

- All models and configurations documented for reproducibility

Repo with full results

Contact Info