Uncovering Fine-Grained Intents in Corporate Email: Methodology, Dataset Characterization, and Label Quality Assessment

h_da
HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES

Patrick Deck

Supervisors: Prof. Dr. Markus Döhring, Prof. Dr. Timo Schürg

Motivation

Corporate email is a cornerstone of workplace communication, crucial for coordination, decision-making, and knowledge sharing. While essential, extracting actionable insights or enabling sophisticated automation is hindered by the complex, unstructured nature of email conversations. Understanding the user's precise goal requires moving beyond broad topics or categorizations.

This research focuses on **fine-grained intents**: the specific, often subtle, communicative goals expressed typically at the sentence level (e.g. distinguishing a request to set up a meeting from a request to reschedule a meeting, or the request of sending a copy of a document from being copied (CC'd) on an email). Capturing this granularity is key for meaningful workflow automation and nuanced communication analysis.

Research Goal

This research aims to answer the following key questions:

- **RQ1 (Methodology & LLMs)**: How can fine-grained user intents be effectively identified, categorized, and labeled at scale within a corporate email corpus, and what role can LLMs play in facilitating this complex task?
- **RQ2 (Dataset Characteristics)**: What are the **structural characteristics** (distribution, semantic separability, lexical patterns of a labeled dataset representing fine-grained corporate email intents, and how do these compare to established **intent benchmark datasets**?)
- **RQ3 (Label Quality & Domain Challenges)**: Given the inherent challenges of semantic ambiguity and potential multi-intent expressions in corporate email, what systematic methods can assess **label quality and consistency**, and what **fundamental domain challenges** do these assessments reveal?

Methodology

- Data Source & Preprocessing:
  - Utilized the large-scale Avocado Research Email Collection [1].
  - Applied email parsing, sentence splitting (spaCy), and initial filtering (e.g., removing ads/spam, selecting relevant sentences).
  - Focus on sentence-level intents to mitigate multi-intent problems by facilitating the problem while deliberately discarding contextual information.
- Targeted Sentence Filtering (Quality Focus):
  - **Rule-Based**: Selected sentences matching request patterns (e.g., "can you", "please") and length constraints (5-15 words).
  - **LLM-Based Quality Scoring (LLaMA 3 8B Instruct)**. Scored candidate sentences on *Intent Clarity*, *Self-Containment*, and *Specificity*. Filtered for high-quality (score 5/5/5) utterances (~19.5k sentences).
- LLM-Powered Feature Generation (Semantic Enrichment):
  - For each high-quality utterance, prompted LLaMA 3 8B Instruct to generate:
    - `explicit_intent` descriptor (generating a descriptor focusing on the *explicit* intent behind an utterance)
    - `implicit_intent` descriptor (generating a descriptor focusing on the *implicit* intent behind an utterance)
    - `purpose_summarization` (generating a brief sentence-level summary)
  - Concatenated embeddings (SBERT: `all-MiniLM-L6-v2`) of these three features to create a rich semantic representation.
- Initial Clustering & Taxonomy Seeding:
  - Applied semantic Agglomerative clustering to the concatenated feature embeddings.
  - **Human-in-the-Loop (HITL)**: Manually inspected clusters to identify coherent intent categories and define the seed taxonomy (54 classes).
- Iterative Dataset Expansion & Refinement (Alternating Phases):
  - **Phase 1 (Expand Known)**: Used supervised methods (SetFit [2], Adaptive Decision Boundary (ADB) [3]) trained on current labels, Cosine Similarity search, and rule-based matching to find new candidate utterances for existing intents. HITL verified candidates.
  - **Phase 2 (Discover New)**: Re-clustering applied to the *remaining* unlabeled data using the LLM features to surface potentially new intent categories. HITL validated new intent categories and added them to the taxonomy/dataset.

Methodology

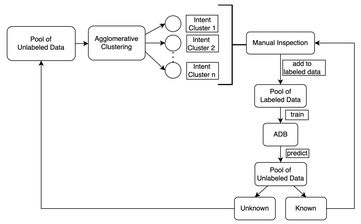
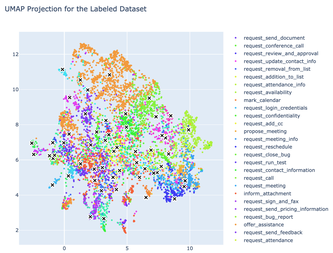


Figure 1. Proposed Approach: To retrieve intent clusters, the unlabeled data is first clustered based on annotations generated by the LLM. After manual inspection using a human-in-the-loop process, an initial set of labeled data is derived. This labeled data is then used to train the ADB Open Intent Classifier, enabling the identification of additional samples for known classes. As more data is added to the labeled set, the pool of unlabeled data is gradually reduced. Clustering is then reapplied to the remaining unlabeled data to repeat the process of discovering new intent categories, as well as adding more diverse examples to existing categories, further expanding the dataset and ensuring a diverse set of samples for each intent category.

Results

Our semi-automated workflow yielded a dataset of **6,785 utterances** across **54 fine-grained corporate email intent classes**. Analysis revealed unique structural properties and challenges:

- Dataset Characteristics (RQ2):
  - UMAP visualization of the embeddings produced for the labeled dataset:



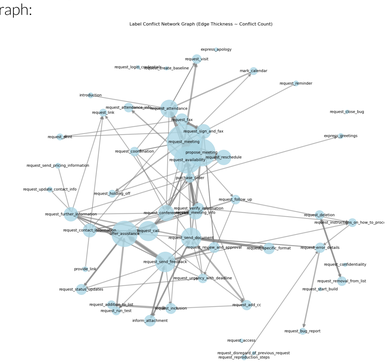
  - **High Semantic Overlap**: Significant overlap observed between classes (Low Silhouette Score: **0.057**, High DBI: **3.257**). Many intents expressed using **shared vocabulary**.
  - **Skewed Distribution**: Final dataset dominated by common operational intents (e.g., `offer_assistance`), while most defined intents are **inherently sparse** in the original corpus (confirmed via random sampling).
- Benchmark Comparison (RQ2):
  - Comparison of the intent benchmark datasets with the labeled dataset:

Dataset	# Intents	Avg. Class Intra Sim	Avg. Inter Sim	Silhouette Score	DBI
SNIPS	7	0.240	0.101	0.151	2.698
BANKING77	77	0.332	0.206	0.156	2.470
CLINC150	150	0.251	0.082	0.220	2.259
StackOverflow	20	0.077	0.015	0.129	2.202
Ours	54	0.265	0.172	0.057	3.257

  - **Structural Similarity**: Closest in complexity (high overlap, reliance on shared vocabulary) to **BANKING77**.
  - **Distinction**: Differs from **CLINC150** (more topical diversity aids separation) and **StackOverflow** (more reliant on distinct keywords).
  - **Challenge**: Positioned as a demanding benchmark for nuanced semantic understanding due to granularity, overlap, and sparsity.

Results

- Label Quality & Domain Challenges (RQ3):
  - Label Conflicts Network Graph:



  - **Cleanlab Findings**: Flagged **~4.2%** of labels as potential inconsistencies.
  - **Nature of Issues**: Inspection revealed conflicts often stem from:
    - **Multi-Intent Utterances**: Single sentences conveying multiple valid intents (e.g., `request_send_document` + `request_add_cc`).
    - **Fuzzy Boundaries**: High confusion between semantically close categories (e.g., `request_meeting` vs. `propose_meeting`).
  - **Implication**: Highlights inherent difficulty of single-label, sentence-level classification for this domain, even after filtering.

Conclusion

- **Developed a Novel Workflow (RQ1)**: Demonstrated a viable semi-automated approach using LLMs (as weak supervisors/feature generators), clustering, and iterative refinement for discovering and labeling fine-grained intents at scale. *LLMs are facilitators, but require human oversight.*
- **Created & Characterized a Challenging Dataset (RQ2)**: Produced a new labeled dataset (6.8k utterances, 54 classes) exhibiting **high semantic overlap**, **skewed distributions**, and **inherent sparsity** of real-world email, positioning it structurally new BANKING77 as a demanding benchmark. *The dataset reflects realistic challenges but is not exhaustive.*
- **Diagnosed Domain Challenges via Label Quality (RQ3)**: Cleanlab analysis (~4.2% potential issues) primarily highlighted **fundamental domain difficulties**, namely the prevalence of **multi-intent utterances** and **fuzzy category boundaries**, even at the sentence level. *This underscores the limitations of single-label approaches for nuanced email understanding.*

Overall Message: Discovering fine-grained intents in corporate email is feasible with LLM-assisted workflows but remains challenging due to inherent semantic complexity and multi-intent utterances. Standard single-label sentence classification is often insufficient.

References

[1] Oard Douglas, William Webber, David Kirsch, and Sergey Golitsynskiy. Avocado research email collection ldc2015t03, 2015.

[2] Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. Efficient few-shot learning without prompts, 2022.

[3] Hanlei Zhang, Hua Xu, and Ting-En Lin. Deep open intent classification with adaptive decision boundary. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14374–14382, May 2021.

Patrick.Deck@stud.h-da.de