ABSTRACT

Understanding fine-grained user intent within corporate email is crucial for workplace automation but faces significant Natural Language Processing (NLP) challenges due to unstructured conversational context, unknown intent taxonomies, and pervasive semantic overlap between related goals. This thesis addresses this complexity by developing, implementing, and evaluating a novel semi-automated workflow for discovering and labeling intents in the large-scale Avocado email corpus. The methodology integrates targeted filtering, weak supervision utilizing Large Language Models (LLMs) (using LLaMA 3 8B Instruct for quality scoring and feature generation), semantic clustering (with SBERT embeddings), and an iterative refinement process. This produced a new labeled dataset of 6,785 utterances across 54 identified fine-grained corporate intent classes.

Characterization of this dataset revealed significant structural properties: a highly skewed distribution (with most defined intents being sparse in the underlying corpus, confirmed via random sampling) and substantial semantic overlap between related classes (low Silhouette Score: 0.057, high DBI: 3.257). Comparative analysis positioned its structural complexity nearest the BANKING77 benchmark, highlighting challenges related to shared vocabulary for nuanced intents, distinct from benchmarks benefiting from greater topical diversity or strong keywords.

Label quality assessment using Confident Learning (Cleanlab) flagged 4.2% of labels as potential inconsistencies. Further analysis indicated these flagged instances primarily highlight the inherent difficulties of assigning definitive single labels in this domain, frequently occurring where utterances reflect semantic ambiguity or fall on inherently fuzzy boundaries between subtle intent categories. This suggests the identified inconsistencies reflect the complexity of mapping nuanced communication onto a discrete taxonomy.

In conclusion, this work contributes: (1) An adaptable workflow demonstrating LLM utility and limitations as weak supervisors in intent discovery; (2) A unique, characterized dataset embodying real-world email complexities like overlap and ambiguity; and (3) Insights from label quality assessment diagnosing inconsistencies tied to inherent domain characteristics. A key finding underscored by this assessment is that even when attempting to simplify the problem by focusing on individual sentences, the prevalence of multi-intent expressions often makes a single label insufficient for accurately capturing the full communicative meaning. The findings collectively emphasize the need for context-aware, potentially multi-label approaches to achieve deeper communication understanding in complex domains like corporate email.

keywords: Intent Discovery, Corporate Email, Fine-Grained Intents, Natural Language Processing, Dataset Creation, Labeling Workflow, Large Language Models

ZUSAMMENFASSUNG

Das Verstehen von feingranularen Nutzerintentionen innerhalb von Unternehmens-E-Mails ist für die Automatisierung des Arbeitsumfelds von entscheidender Bedeutung, stellt jedoch aufgrund des unstrukturierten Konversationskontexts, unbekannter Intentionstaxonomien und weitreichender semantischer Überlappungen zwischen verwandten Zielen eine erhebliche Herausforderung für die natürliche Sprachverarbeitung (NLP) dar. Diese Arbeit befasst sich mit dieser Komplexität durch die Entwicklung, Implementierung und Evaluierung eines neuartigen semi-automatischen Workflows zur Erkennung und Labeling von Intentionen im umfangreichen Avocado E-Mail-Korpus. Die Methodik integriert gezielte Filterung, schwache Überwachung unter Verwendung von Large Language Models (LLMs) (unter Verwendung von LLaMA 3 8B Instruct für Qualitätsbewertung und Feature-Generierung), semantisches Clustering (mit SBERT Embeddings) und einen iterativen Verfeinerungsprozess. Auf diese Weise entstand ein neuer gelabelter Datensatz mit 6.785 Äußerungen über 54 identifizierte, feingranulare Klassen von Intentionen aus dem Unternehmensumfeld.

Die Charakterisierung dieses Datensatzes offenbarte signifikante strukturelle Eigenschaften: eine stark verzerrte Verteilung (im zugrunde liegenden Korpus sind die meisten definierten Intentionen nur selten zu finden, was durch Stichproben bestätigt wurde) und erhebliche semantische Überlappungen zwischen verwandten Klassen (niedriger Silhouette Score: 0,057; hoher DBI: 3,257). Eine vergleichende Analyse ergab, dass die Struktur am nächsten an dem BANKING77-Benchmark liegt, was die Herausforderungen im Zusammenhang mit dem geteilten Vokabular für nuancierte Intentionen hervorhebt, im Gegensatz zu Benchmarks, die von einer größeren thematischen Vielfalt oder starken Schlagwörtern profitieren.

Bei der Bewertung der Label-Qualität mit Confident Learning (Cleanlab) wurden 4,2% der Labels als potenzielle Unstimmigkeiten eingestuft. Weitere Analysen ergaben, dass diese markierten Instanzen in erster Linie die inhärenten Schwierigkeiten bei der Zuweisung definitiver Single-Labels in diesem Bereich verdeutlichen. Sie treten häufig dort auf, wo Äußerungen semantische Mehrdeutigkeit widerspiegeln oder auf unscharfe Grenzen zwischen subtilen Intentionskategorien fallen. Dies deutet darauf hin, dass die festgestellten Unstimmigkeiten die Komplexität der Zuordnung nuancierter Kommunikation zu einer diskreten Taxonomie widerspiegeln.

Zusammenfassend trägt diese Arbeit zu Folgendem bei: (1) Einem anpassungsfähigen Workflow, der den Nutzen und die Grenzen von LLMs als schwache Supervisoren bei der Entdeckung von Intentionen aufzeigt; (2) einem einzigartigen, charakterisierten Datensatz, der die Komplexität von E-Mails in der realen Welt verkörpert, wie z.B. Überlappungen und Mehrdeutigkeit; und (3) Erkenntnissen aus der Bewertung der Label-Qualität, die Unstimmigkeiten in Verbindung mit inhärenten Domänenmerkmalen diagnostiziert. Eine wichtige Erkenntnis, die durch diese Auswertung unterstrichen wird, ist, dass selbst bei dem Versuch, das Problem zu vereinfachen, indem der Fokus

auf einzelne Sätze gelegt wurde, die Prävalenz von Ausdrücken mit mehreren Intentionen oft dazu führt, dass ein einzelnes Label nicht ausreicht, um die ganzheitliche kommunikative Bedeutung exakt zu erfassen. Die Ergebnisse unterstreichen den Bedarf an kontextbezogenen, potenziell Multi-Label-Ansätzen, um ein tieferes Verständnis der Kommunikation in komplexen Domänen wie Firmen-E-Mails zu erreichen.

Schlagwörter: Intent Discovery, Corporate Email, Fine-Grained Intents, Natural Language Processing, Dataset Creation, Labeling Workflow, Large Language Models