# DIDECO: An Annotated Dataset for Intent Detection in Digital Communications

**Senaid Popovic**[1,2], **Damien Riquet**[1], **Maxime Meyer**[1], **Fabien Lauer**[2], **Yannick Parmentier**[2]
[1]Hornetsecurity, Lille, France
[2]Université de Lorraine, CNRS, LORIA, Nancy, France
[1] firstname.lastname@vadesecure.com, [2] firstname.lastname@loria.fr

**Abstract**

This paper presents DIDECO, the first annotated dataset specifically designed for detecting both explicit and implicit intents in digital communications. We address a critical gap in cybersecurity research by developing a comprehensive taxonomy that distinguishes between explicit communicative goals (what is requested) and implicit persuasion mechanisms (how compliance is engineered). Grounded in Speech Act Theory and persuasion psychology principles, our taxonomy encompasses 20 distinct intent categories across explicit and implicit intents. We annotated 220 LLM-generated spear-phishing emails using a multi-label protocol with six trained annotators, yielding 2,162 intent annotations that reveal the layered complexity of malicious communications. Our analysis demonstrates that sophisticated attacks employ multiple intents, combining explicit communicative goals with implicit persuasion strategies. This dataset provides resources for developing intent-aware detection systems capable of identifying sophisticated social engineering attacks through semantic analysis.

**Keywords:** intent detection, dataset, digital communications, communicative intent, annotation methodology, cybersecurity, phishing, social engineering

## 1. Introduction

The analysis of intent in digital communications is increasingly critical as organizations rely on digital channels for internal and external interactions. The threat landscape has evolved significantly, with attackers becoming more sophisticated in creating textual threats that can evade traditional detection systems. Modern malicious actors leverage advanced social engineering techniques, AI-generated content, and persuasion strategies that make their communications appear legitimate while carrying harmful intent. The cost of misinterpreting intent is substantial, with successful phishing attacks alone costing businesses billions of dollars annually. The FBI Internet Crime Report documented over $10.3 billion in reported losses, with Business Email Compromise (BEC) schemes accounting for more than $2.9 billion (Federal Bureau of Investigation, 2023).

As attackers become more adept at mimicking legitimate communication and exploiting psychological manipulation techniques, traditional keyword-based and signature-based detection methods prove increasingly inadequate. This evolution necessitates a fundamental shift toward deeper semantic analysis and intent understanding in cybersecurity systems. To effectively counter these threats, security systems must extract comprehensive semantic information from textual communications, analyzing not only what is explicitly stated but also the implicit persuasion strategies and psychological manipulation techniques employed. Despite this critical need, the development of effective automatic intent detection systems is slowed by the lack of large-scale, comprehensively annotated datasets that capture the full spectrum of communicative intentions in digital environments.

This paper addresses these limitations through the construction of a large-scale annotated dataset based on the development of an explicit/implicit intent taxonomy and a LLM-based annotation pipeline. Our specific objectives are:

1. **Building a comprehensive taxonomy**: Development of a classification framework that distinguishes between explicit and implicit intents in digital communications, providing clear definitions and boundaries for consistent annotation.

2. **Creating a dataset**: Creation of the first comprehensive dataset for explicit and implicit intent detection, containing annotated examples across multiple intent categories.

3. **Validating our taxonomy**: Empirical analysis of label distributions across 220 annotated emails, demonstrating the applicability and coverage of the taxonomy across spear-phishing intent categories.

4. **Provisioning the resource**: Delivery of a dataset and annotation methodology that enables the research community to advance intent detection systems and develop more effective cybersecurity applications.

https://github.com/Sneusneu/Dideco/



*Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 6808–6822
11–16 May 2026. 

The paper is organized as follows. Section 2 reviews related work on theoretical foundations, email intent detection, and existing datasets. Section 3 presents our explicit/implicit intent taxonomy grounded in Speech Act Theory and persuasion psychology, followed by Section 4 which describes the corpus design and annotation protocol. Section 5 analyzes the resulting dataset distribution patterns and discusses implications for detection systems, before Section 6 outlines future applications and extensions.

## 2. Related Work

### 2.1. Theoretical Foundations of Intent in Language

**Rhetorical and Psycholinguistic Perspectives.** The study of intent in language has deep roots in rhetoric, discourse analysis, and psycholinguistics, where communication is viewed as goal-oriented action shaped by audience design, credibility, emotion, and argumentative structure. Recent computational work on online persuasion highlights two complementary lenses: linguistic analysis focusing on stylistic and rhetorical cues, and argumentative approaches modeling broader discourse structure and stance, often trading interpretability for predictive power (Bassi et al., 2024). This body of work motivates our decision to separate explicit intents (what is being done with words) from implicit intents (how influence is exerted), and to ground the latter in influence mechanisms rather than surface keywords alone.

A central challenge is that implicit intent is frequently expressed through indirectness, presuppositions, hedging, and other pragmatics-driven phenomena whose recovery requires pragmatic inference over shared context and social norms (Frank and Goodman, 2012; Bassi et al., 2024). This aligns with recent computational accounts showing that implied meaning is systematically harder to model and annotate than explicitly stated content, and it helps explain why annotators tend to exhibit lower confidence and higher disagreement on implicit intent labels than on explicit speech acts (Pavlick and Kwiatkowski, 2019; Uma et al., 2021).

Finally, the boundary between "legitimate persuasion" and "harmful manipulation" is partly normative and context-dependent. Rather than claiming a universal philosophical criterion, we operationalize intent categories for the specific setting of security-relevant digital communication (e.g., spearphishing), and we document annotation guidelines to make the criteria explicit and auditable (Bassi et al., 2024).

**Speech Act Theory and Illocutionary Force.** The theoretical foundations of our taxonomy are rooted in Speech Act Theory, introduced by (Austin, 1962), which proposed that language serves not only to describe the world but also to perform actions. (Searle, 1969) expanded this work by developing the concept of indirect speech acts and proposing a taxonomy of illocutionary acts including assertives, directives, commissives, expressives, and declaratives. This classification has had lasting impact on intent detection research, with applications to email classification (Cohen et al., 2004) and social media analysis (Carr and Hayes, 2017). The theory's emphasis on distinguishing between what is literally said and what is intended provides a crucial lens for understanding the difference between explicit and implicit intents in digital communications.

**Persuasion Psychology and Social Engineering.** The implicit intent categories in our taxonomy draw heavily from Cialdini's principles of persuasion (Cialdini, 1993), which identify authority, social proof, scarcity, and reciprocity as fundamental mechanisms of influence. We also incorporate insights from (Gragg, 2003) on social engineering techniques and (Stajano and Wilson, 2011) on scam principles.

**Persuasion and manipulation technique detection in NLP.** Beyond cybersecurity, NLP has operationalized fine-grained persuasion/manipulation strategies through shared tasks and annotated corpora, particularly in propaganda technique detection, where spans are labeled with specific influence tactics (Da San Martino et al., 2019, 2020). While these resources target news and political discourse rather than email, they demonstrate that influence can be modeled as a multi-label inventory of tactics, complementing our implicit-intent taxonomy for social engineering settings.

**Cybersecurity and Digital Communication Analysis.** Recent research (Ferreira and Teles, 2019) on persuasion in phishing emails has demonstrated the importance of understanding both explicit requests and implicit persuasion techniques in cybersecurity contexts. Our framework extends this work by providing a comprehensive categorization applicable to both legitimate and malicious communications.

### 2.2. Email Intent Detection

**Email Intent Detection and Classification.** Our taxonomy development builds upon established research in organizational email usage, particularly (Dabbish et al., 2005) on understanding email use

and predicting action on messages. (Cohen et al., 2004) applied Speech Act Theory to email classification, while (Carvalho and Cohen, 2005) introduced dependency-network-based collective classification showing significant improvements over baseline classifiers. A critical advancement came with (Lampert et al., 2010), who introduced email zoning techniques, demonstrating that segmenting emails into functional zones (new content, quoted replies, signatures) significantly improves intent classification. (Sappelli et al., 2016) further refined methods for assessing email intent, while (Yang et al., 2017) provided insights on enterprise email reply behavior. The shift to deep learning brought substantial improvements: (Wang et al., 2019) introduced the Dynamic-Context Recurrent Neural Network (DCRNN) model, which leverages both target sentences and full email context, dynamically integrating contextual information to enhance intent detection accuracy in complex email threads.

**Evolution of Email-Based Threats.** The threat landscape has evolved dramatically with AI-assisted attack generation. (Nahmias et al., 2024) introduced a dataset of 333 spear-phishing emails generated using LLM-based agents and knowledge graph technologies, representing a critical shift in understanding modern threats. Attackers can now generate highly personalized, context-aware phishing emails at scale, making them increasingly difficult to distinguish from legitimate communications. Traditional detection methods based on keywords or simple patterns are inadequate against these sophisticated, AI-generated threats. This synthetic dataset enables researchers to evaluate detection methods against advanced threats not adequately represented in traditional datasets.

### 2.3. Datasets for Intent Detection Research

**General Intent Detection Benchmarks.** The Natural Language Understanding (NLU) research community has developed several benchmark datasets. The ATIS (Airline Travel Information System) dataset (Hemphill et al., 1990), containing 5,871 utterances with 17 intents, has been widely used despite its age and narrow domain. (Coucke et al., 2018) introduced SNIPS, a more diverse dataset covering seven domains with 14,484 utterances. (Casanueva et al., 2020) introduced BANKING77, with 13,083 customer service queries labeled with 77 fine-grained intents, providing a challenging benchmark due to its granular intent distinctions.

**Email-Specific Datasets.** Email intent detection has primarily relied on two corpora. The Enron email corpus (Klimt and Yang, 2004), containing over 500,000 messages from approximately 150 users, has been most commonly used despite privacy concerns and being outdated (early 2000s). The Avocado dataset (Oard et al., 2015), with over 900,000 messages from 279 accounts, provides richer metadata but still represents a single company's communication style. Both face significant limitations: temporal obsolescence, lack of comprehensive intent annotations, privacy concerns, and limited domain generalizability.

**Limitations and Research Gaps.** Current datasets exhibit several critical limitations. First, lack of comprehensive annotation for both explicit and implicit intents—existing datasets typically focus on one or the other. Second, most are domain-specific or language-specific, limiting cross-domain research. Third, synthetic or recent data reflecting modern communication patterns and AI-generated content is scarce. Fourth, few datasets are designed for cybersecurity applications requiring distinction between legitimate and malicious intent. Finally, insufficient representation of implicit persuasion techniques characterizing modern social engineering attacks.

**LLM-assisted annotation and weak supervision.** Recent work has explored using large language models as annotators to scale labeling via "silver" annotations, often paired with human oversight, calibration, and audits to manage noise and bias (Tan et al., 2024). This paradigm is particularly relevant when labels require pragmatic inference (as with implicit intents), where fully manual annotation can be slow and inconsistent. In our setting, LLM-assisted annotation is positioned as a complement to human labeling: we prioritize high-quality human annotations where feasible and use automatic labels to expand coverage, while reporting current statistics and acknowledging limitations.

Our work addresses these gaps by creating a large-scale dataset with comprehensive annotations covering both explicit and implicit intents across multiple communication types, with particular attention to cybersecurity-relevant categories.

## 3. Taxonomy of Intents

Our global approach builds upon a theoretically grounded taxonomy, which uses a hierarchical architecture that enables analysis at multiple granularity levels. The complete taxonomy is available in Appendix A. At the top level of this hierarchy, the taxonomy first distinguishes between explicit intents and implicit intents.

### 3.1. Research Gap and Contribution

The research field addressing textual (and specifically emails) social engineering has developed along two distinct trajectories. Psychological frameworks predominantly examine the persuasion mechanisms employed by attackers, analyzing principles such as authority, scarcity, and social proof that influence victim decision-making (Cialdini, 1993; Stajano and Wilson, 2011; Ferreira and Teles, 2019). These frameworks provide deep insights into how manipulation operates but typically do not categorize the explicit objectives that attackers pursue through these psychological strategies. In parallel, technical email classification systems and communication-focused taxonomies have concentrated on identifying surface-level requests and actions, such as financial transfers, credential gathering, and information access (Dabbish et al., 2005; Gragg, 2003; Bennett and Carbonell, 2005). While valuable for applied detection purposes, these approaches generally treat psychological manipulation as secondary or implicit, without systematically integrating it into their analytical structures. This separation creates a gap in understanding complex attacks. Real-world threats rarely rely on psychological manipulation alone or explicit requests in isolation. Consider a spear-phishing email impersonating a trusted colleague (identity establishment) requesting credential verification (call to action) while invoking shared organizational responsibility (social proof) and creating time pressure (temporal scarcity). Such an attack exploits multiple intent types simultaneously across both psychological and communicative dimensions. Existing frameworks must choose whether to classify this attack primarily by its persuasion mechanism, its requested action, or its impersonation tactic, losing sight of the sophistication that makes the attack effective. Our taxonomy tries to address this gap through a combination of both explicit and implicit intents. By separating explicit intents (what is requested) from implicit intents (how compliance is engineered), we enable analysis of both dimensions simultaneously and their interactions. By bridging communicative and psychological dimensions while maintaining actionability across multiple security contexts, our taxonomy provides the conceptual infrastructure for systematic email threat analysis.

### 3.2. Explicit Intents

Explicit intents in digital communications are clearly expressed objectives that aim to influence the recipient's behavior or obtain specific results. Based on speech act theory (Searle, 1969), we propose four categories that cover the main objectives of explicit digital communications.

**Information Manipulation.** This fundamental category reflects the central role of information exchange and control in digital interactions, a pattern documented extensively in research on both professional communications and cybersecurity threats (Dabbish et al., 2005; Ferreira and Teles, 2019; Gragg, 2003). The subcategory **Information Gathering** includes actions aimed at obtaining data, ranging from legitimate verifications to malicious attempts to harvest credentials, personal details, or organizational intelligence. **Information Control** concerns attempts to manage information flow or system access, such as requests to modify permissions, update system settings, or alter security configurations. This distinction is supported by phishing research showing that attacks often begin by collecting information before attempting to manipulate systems or extract resources.

**Resource Acquisition.** This category emerges from analyses of both legitimate business transactions and fraudulent schemes (Dabbish and Kraut, 2006; Stajano and Wilson, 2011), recognizing that resource requests constitute a distinct communicative objective beyond simple information exchange. We distinguish **Direct Financial Gain**, which involves immediate money transfers or requests for monetary information such as bank account details or payment credentials, and **Indirect Resource Acquisition**, which includes requests for purchases of gift cards, services, or asset transfers that provide value without direct monetary exchange. Research on business email compromise confirms this distinction, documenting how attackers employ both direct financial requests and more subtle resource manipulations depending on the target and context (Federal Bureau of Investigation, 2023).

**Trust Establishment.** Studies of both legitimate and malicious digital communications demonstrate that establishing credibility represents a distinct communicative goal requiring its own analytical category (Gragg, 2003). The subcategory **Identity Establishment** concerns efforts to verify or assert the sender's identity and establish their authority within organizational or social hierarchies—for instance, claiming to be a senior executive, IT administrator, or trusted colleague. **Legitimacy Creation** focuses on establishing context and credibility through institutional associations, such as references to known organizations like PayPal, Microsoft, or governmental agencies, or mentions of legitimate business processes like compliance requirements or system updates. This distinction recognizes that personal identity and organizational context function as distinct but complementary trust signals, each requiring different verifica-

tion approaches.

**Call to Action.** Drawing on previous work on detecting action elements in communications (Bennett and Carbonell, 2005), we identify the call to action as representing a unique communicative objective, distinct from long-term compliance strategies or relationship-building efforts. **Direct Action Requests** involve explicit instructions for immediate actions, such as "Click this link to reset your password now" or "Wire the funds to this account today," where the expected behavior is clearly specified and temporally immediate. **Indirect Action Facilitation** encompasses preparatory steps that enable or necessitate future actions, such as "Keep these credentials for your next login" or "Save this contact information for upcoming communications," where the current request sets up conditions for subsequent compliance. This distinction captures the temporal dimension of digital communication intents, recognizing that immediate actions are often preceded by preparatory work that facilitates subsequent behavior and creates commitment pathways.

### 3.3. Implicit Intents

Implicit intents exploit psychological and social principles to indirectly influence behavior. Our taxonomy adopts the established classification by Ferreira et al. (Ferreira and Teles, 2019), which synthesizes the foundational work of Cialdini (Cialdini, 1993), Gragg (Gragg, 2003), and Stajano & Wilson (Stajano and Wilson, 2011) into a unified structure validated for analyzing persuasion tactics in digital communications. This approach allows us to benefit from a conceptual framework already established in the literature while integrating it with our explicit intent layer to enable dual-dimension analysis.

**Authority.** This category exploits individuals' conditioned tendency to submit to authority figures or authority signals, leveraging the well-documented compliance effects observed when requests appear to come from legitimate or high-status sources. This manifests through hierarchical language, organizational titles, power dynamics in phrasing, and references to institutional authority, creating pressure to comply without critical evaluation of the request's legitimacy.

**Social Proof.** These intents use individuals' propensity to determine appropriate behavior by observing what others do, exploiting conformity mechanisms and social validation processes. The literature distinguishes three manifestations: **herd mentality**, which encompasses conformity behaviors based on perceived group actions such as "Everyone in the department has already completed this"; **diffusion of responsibility**, characterized by reduced individual accountability in collective contexts through framing like "We all need to do this together"; and **moral duty**, which involves appeals to prosocial behavior and altruism such as "Help us protect other users" or charitable cause exploitation.

**Liking, Similarity, and Deception.** This category groups techniques related to interpersonal attraction and deception mechanisms. Research distinguishes **general deception**, which involves false scenarios or misleading contexts such as fabricated emergencies or fictitious business situations, from **deceptive relationships**, characterized by falsified connections or similarities like false claims of shared colleagues, or professional backgrounds. **Liking and similarity** exploits genuine or claimed shared characteristics and affinities, leveraging the principle that people are more likely to comply with requests from those they perceive as similar to themselves or with whom they feel rapport.

**Distraction.** These techniques divert attention from critical evaluation and integrate cognitive bias exploitation to reduce deliberative thinking. The literature identifies four mechanisms: **scarcity**, encompassing both time-based limitations like "Offer expires in 24 hours" and resource-based constraints such as "Only 3 spots remaining"; **overloading**, which creates cognitive overload through time pressure, information complexity, or simultaneous demands that exceed processing capacity; **strong affect**, involving emotional state manipulation through fear ("Your account will be suspended"), excitement ("You've won"), or anxiety ("Urgent security alert"); and **need and greed**, which exploits fundamental desires and appeals to basic human needs or acquisitive tendencies through promises of financial gain, exclusive opportunities, or solutions to pressing problems.

**Commitment, Integrity, and Reciprocation.** This category encompasses techniques that exploit psychological consistency principles and social exchange norms. The framework includes **integrity**, which exploits the general presumption of honesty in human interactions and the assumption that communications are truthful until proven otherwise; **consistency**, involving appeals to behavioral consistency with past actions or stated values, such as "As you previously indicated your interest in..."; **commitment**, which leverages previous commitments through reminders of past agreements, escalation of prior small requests, or coercive framing that makes refusal psychologically costly; and **reciprocation**, creating obligations through mentions

of favors, services, or benefits provided, activating the felt social obligation to return a favor even when the initial "favor" was unsolicited or fabricated.

# 4. Building the dataset

## 4.1. Corpus Design and Collection

Our evaluation corpus includes a dataset of 220 high-quality spear-phishing emails specifically designed to represent the emerging threat of LLM-generated social engineering attacks (Nahmias et al., 2024). All emails in the corpus are in English. These emails were generated by an automated system that leverages large language models for both target reconnaissance and personalized message creation. This dataset, while relatively small in size, serves as a foundation for human annotation and manual validation of our taxonomy. The limited scope allows for manual annotation to establish ground truth labels and to validate the applicability of our intent categories to spear-phishing content. A natural next step is to scale annotation by applying LLM-based methods to larger email corpora, building on the human-annotated ground truth established here. These larger datasets will incorporate emails from diverse established sources, including the Enron email corpus, Apache SpamAssassin collections, and other publicly available email datasets, enabling comprehensive evaluation of our taxonomy across a broader spectrum of email types and communication contexts.[1]

## 4.2. Annotation Protocol

The following guidelines establish a standardized approach for applying our taxonomy to digital communications. These guidelines ensure consistency and reliability when identifying and categorizing both explicit and implicit intents present in textual communication such as emails.

**Purpose and Scope.** This annotation protocol facilitates the systematic identification of communicative intents using our proposed taxonomy. The primary goal is to enable consistent analysis of digital communications, with particular focus on detecting both legitimate and potentially malicious intents. Each communication, typically an individual email or SMS message, will be analyzed as a complete unit, though specific intent markers may be identified at the sentence or paragraph level within the larger context. When analyzing threaded conversations or document chains, each message or document should be treated as a separate communication unit.

[1] https://github.com/Sneusneu/Dideco/

**Multi-label Annotation.** We employ a multi-label annotation approach, recognizing that most digital communications contain multiple intent categories simultaneously. Annotators should identify all applicable intent categories present in a communication, rather than forcing selection of a single dominant intent. This approach acknowledges the layered nature of persuasive communication, where explicit requests often leverage multiple implicit persuasion techniques simultaneously.

**Annotation Platform.** All annotations were conducted using the POTATO system (Pei et al., 2022), which provides specialized interfaces for hierarchical multi-label classification. The platform allows annotators to document their decision-making process and highlight specific textual evidence for each annotation.

**Annotators.** We employed six annotators from our research team to ensure robust inter-annotator agreement and reliable categorization. This team composition balances expertise with diverse perspectives: two annotators possess prior knowledge of the taxonomy framework and contributed to its development, while the remaining four annotators were recruited specifically for this task and trained using detailed annotation guidelines. This mixed-expertise approach allows us to evaluate both the internal validity of the taxonomy and its usability by non-experts who underwent brief training. This training follows a structured protocol: annotators first reviewed the taxonomy documentation, then took part in a guided practice session with pre-annotated examples, and finally discussed and resolved discrepancies collectively. Inter-annotator agreement is reported in Appendix D.

# 5. Dataset Analysis and Discussion

**Corpus Composition and Scale.** Following the annotation protocol described in Section 4, we annotated 220 emails from our corpus, yielding a comprehensive dataset of 2,162 intent annotations across 20 distinct categories. Through this annotation task, we note that sophisticated email threats rarely employ single-intent strategies; instead, they combine multiple explicit requests with implicit persuasion mechanisms to achieve their communicative goals.

## 5.1. Distribution of Explicit and Implicit Intents

One of the findings from our corpus analysis is the distribution between explicit and implicit intent annotations. As shown in Table 1, explicit intents

account for 686 annotations (31.7%), while implicit intents comprise 1,476 annotations (68.3%).

| Intent Category | Freq. | % |
|---|---|---|
| **Explicit Intents** | | |
| Trust Establishment | 342 | 15.8% |
| Call to Action | 295 | 13.6% |
| Information Manipulation | 39 | 1.8% |
| Resource Acquisition | 10 | 0.5% |
| *Explicit Subtotal* | *686* | *31.7%* |
| **Implicit Intents** | | |
| Commitment & Reciprocation | 470 | 21.7% |
| Social Proof | 340 | 15.7% |
| Distraction | 297 | 13.7% |
| Liking, Similarity & Deception | 211 | 9.8% |
| Authority | 158 | 7.3% |
| *Implicit Subtotal* | *1,476* | *68.3%* |
| **Total** | **2,162** | **100.0%** |

Table 1: Distribution of main intent categories in the DIDECO corpus

Using this table, we can see that *Trust Establishment* emerges as the most frequent explicit category (342 instances, 15.8%), highlighting the critical role of credibility building in malicious communications. *Call to Action* follows with 295 instances (13.6%), reflecting the action-oriented nature of spear-phishing emails. *Information Manipulation* (39 instances, 1.8%) and *Resource Acquisition* (10 instances, 0.5%) round out the explicit categories; their low frequency reflects the source corpus's focus on credential-harvesting attacks rather than direct financial fraud, as discussed below.

Among implicit intents, *Commitment & Reciprocation* is the most prevalent category (470 instances, 21.7%), demonstrating the widespread use of consistency principles and reciprocity norms in persuasive communication. *Social Proof* follows with 340 instances (15.7%), reflecting frequent invocation of collective legitimacy and peer behavior even in highly targeted one-to-one attacks. *Distraction* techniques, including scarcity appeals and emotional manipulation, account for 297 instances (13.7%), underlining their importance in bypassing critical evaluation. *Liking, Similarity & Deception* strategies appear in 211 instances (9.8%), while *Authority* signals occur in 158 instances (7.3%).

## 5.2. Discussion

The distribution patterns revealed in our analysis illuminate several important characteristics of persuasive communication in digital environments. The predominance of implicit intents (68.3%) over explicit intents (31.7%) confirms the relevance of our dual-taxonomy approach and underscores that systems designed to detect malicious communications must analyze implicit persuasion mechanisms with at least as much priority as explicit requests to achieve comprehensive threat detection.

The prominence of *Trust Establishment* as the most frequent explicit intent reflects a fundamental aspect of spear-phishing communication: attackers invest heavily in credibility-building before issuing any explicit request, particularly when impersonating colleagues or organizational figures. The high frequency of *Call to Action* (13.6%) confirms that most spear-phishing emails are ultimately oriented toward eliciting specific behaviors, whether clicking links, providing credentials, or approving actions.

The prominence of *Social Proof* (340 instances, 15.7%) as the second most frequent implicit category is noteworthy. Rather than being reserved for mass-scale phishing campaigns, social proof mechanisms — such as references to shared group membership, organizational norms, and peer behavior — appear frequently even in highly targeted one-to-one attacks, suggesting that attackers routinely leverage collective legitimacy as a persuasion vector regardless of targeting specificity.

**Near-absence of Resource Acquisition and Information Manipulation.** As shown in Table 1, *Resource Acquisition* (0.5%) and *Information Manipulation* (1.8%) are the rarest explicit categories by a substantial margin. This is attributable to the nature of the source corpus (Nahmias et al., 2024), which was generated to simulate credential-harvesting and link-clicking spear-phishing rather than Business Email Compromise (BEC) or financial fraud attacks. Emails designed to harvest credentials do not explicitly request financial transfers or directly manipulate information channels, which are the scenarios that most reliably trigger these two labels. This finding highlights a structural limitation of the current corpus noted in Section 5.2: it does not represent the full spectrum of spear-phishing subtypes. We recommend that future corpus extensions explicitly include BEC and financial fraud scenarios to ensure balanced coverage across all explicit intent categories.

The multi-label nature of our annotations reveals that sophisticated communications typically employ multiple intent categories simultaneously. Our analysis shows that emails containing explicit requests almost invariably incorporate implicit persuasion mechanisms. This layering effect suggests that attackers understand the importance of combining multiple influence vectors to increase success rates. For instance, a typical spear-phishing email might establish authority, create urgency through scarcity appeals, and invoke reciprocity norms while explicitly requesting credential information. This finding has implications for detection systems, which must be capable of identifying and analyzing multiple concurrent intents rather than treating communica-

tions as single-intent instances.

The distribution patterns also highlight potential vulnerabilities in current detection approaches. The high prevalence of *Distraction* techniques (13.7%) suggests that attackers successfully exploit cognitive biases by overwhelming recipients with emotional appeals, time pressure, and arousal tactics. Detection systems that focus on explicit content will miss these subtle influence mechanisms, which together with *Commitment & Reciprocation* and *Social Proof* account for over half of all annotated labels.

One notable observation concerns the relationship between intent categories and detection difficulty. Our human annotators reported highest confidence when identifying explicit intents like *Call to Action* and *Trust Establishment*, while implicit categories such as *Liking, Similarity & Deception* and *Commitment & Reciprocation* required more careful analysis and deliberation. This variation in identification difficulty suggests that automated detection systems may require different architectural approaches or training strategies for explicit versus implicit intent categories.

**Limitations.** Several limitations of the current dataset should be acknowledged. First, the 220 annotated emails are drawn from a single synthetic corpus (Nahmias et al., 2024) generated using a specific LLM-based pipeline targeting particular spear-phishing scenarios. While this provides a controlled and high-quality annotation foundation, it limits diversity: the emails are stylistically homogeneous and represent only a subset of real-world spear-phishing strategies. The corpus is thus representative of certain types of spear-phishing (e.g., credential harvesting, urgency-driven requests) but not of the full spectrum of attack types documented in the literature (Nahmias et al., 2024). Second, the corpus contains no legitimate communications, which restricts the dataset's applicability to binary malicious/benign classification tasks and may introduce detection biases in downstream models. Third, as a solely English-language resource, DIDECO does not yet address the multilingual dimension of social engineering threats. These limitations are recognized starting points for the extensions described in the Future Work section below.

## 6. Applications and Future Work

The DIDECO dataset and our explicit/implicit intent taxonomy enable multiple research directions and practical applications in cybersecurity, natural language processing, and organizational communication analysis. This section outlines immediate applications and future extensions that can build upon our foundational work.

**Intent Detection System Development.** The most direct application of our dataset is enhancing threat detection models by providing them with additional detection signals. Our multi-label annotations provide ground truth for developing classifiers that can identify both explicit requests and implicit persuasion mechanisms.

Beyond binary malicious/benign classification, intent-aware systems can provide contextual risk assessments that account for the specific combination of persuasion tactics employed. For instance, an email exhibiting high authority signals combined with urgency-inducing distraction techniques might warrant elevated scrutiny even if no malicious indicators are present in content alone. This nuanced approach addresses the limitations of traditional detection methods that rely primarily on surface-level features or known attack signatures.

Our taxonomy also supports explainable AI approaches in cybersecurity. Rather than providing opaque risk scores, intent-based detection systems can explain their assessments in terms of identified persuasion mechanisms and explicit requests. Security analysts can receive alerts indicating specific intent categories detected, enabling more informed decision-making about threat response and prioritization.

**Dataset Extension and Scaling.** While our initial corpus focuses on LLM-generated spear-phishing emails, the taxonomy and annotation methodology generalize to diverse communication contexts. Future work will create larger datasets and adapt the annotation methodology to scale beyond the current 220-email subset. In particular, we will apply LLM-based annotation to larger email corpora once its agreement with human annotations is validated, and we will incorporate emails from established sources (e.g., Enron, Apache SpamAssassin) to improve coverage and generalizability.

A priority is to incorporate legitimate business communications from the Enron corpus (Klimt and Yang, 2004) and other organizational email datasets. Deployable systems must distinguish malicious manipulation from legitimate business persuasion; the current corpus focuses on malicious content, so adding benign examples is necessary to reduce false positives and to train intent-aware classifiers that operate in mixed settings.

We also plan to expand data sources beyond spear-phishing to include other attack vectors such as SMS phishing, social media direct messages, and instant messaging platforms where social engineering attacks increasingly occur.

**Automatic Annotation and Model Adaptation.** Human annotators reported lower confidence and more deliberation for implicit intent categories (e.g., Liking/Similarity/Deception, Commitment & Reciprocation), which suggests that training automated models on these categories will be more challenging. Future work will therefore focus on adapting strategies for automatic annotation by LLMs: refining prompts, few-shot setups, and consistency checks so that scalable annotation remains aligned with human judgments. This direction is essential to extend the dataset to larger corpora while preserving annotation quality.

**Multilingual Intent Detection.** Our current dataset focuses exclusively on English-language communications, but social engineering attacks are a global phenomenon requiring multilingual detection capabilities. Extending our taxonomy and annotation methodology to other languages presents both opportunities and challenges that merit careful consideration.

## 7. Conclusion

This paper presents DIDECO, the first annotated dataset specifically designed for explicit and implicit intent detection in digital communications. Our work addresses a critical gap in cybersecurity research by providing a comprehensive resource for developing and evaluating intent-aware detection systems capable of identifying sophisticated social engineering attacks.

Our primary contributions include the development of a novel taxonomy that distinguishes between explicit communicative goals and implicit persuasion mechanisms, grounded in Speech Act Theory and established principles from persuasion psychology. This theoretical foundation provides conceptual boundaries while maintaining applicability to real-world digital communications. The taxonomy encompasses 20 distinct intent categories organized into explicit intents such as Trust Establishment, Call to Action, Information Manipulation, and Resource Acquisition, alongside implicit categories including Authority, Social Proof, Commitment and Reciprocation, Distraction, and Liking, Similarity and Deception.

The resulting dataset of 220 annotated emails, comprising 2,162 intent annotations, reveals important patterns in how explicit requests and implicit persuasion mechanisms combine in digital communications. The predominance of implicit intents (68.3%) over explicit intents (31.7%) validates our dual-taxonomy approach and highlights that sophisticated spear-phishing attacks rely primarily on psychological influence mechanisms rather than overt requests, while the multi-label nature of annotations reflects the layered complexity of persuasive communication.

Looking forward, the DIDECO dataset provides a foundation for multiple research directions. Immediate applications include training supervised learning models for intent-aware threat detection, developing explainable AI systems that can articulate the specific persuasion mechanisms present in communications, and conducting comparative analyses of legitimate versus malicious persuasion tactics. Extensions to broader communication contexts, additional attack vectors, multilingual settings, and longitudinal analysis will further enhance the dataset's utility.

Beyond technical contributions, this work advances the broader goal of understanding and defending against social engineering in digital environments. As attackers increasingly leverage AI to generate sophisticated, personalized threats, defensive capabilities must evolve beyond surface-level pattern matching toward deeper semantic analysis and intent understanding. Our taxonomy and dataset provide resources for this evolution, enabling the research community to develop more effective, explainable, and robust detection systems.

## Ethics Statement

The DIDECO dataset is built from LLM-generated synthetic spear-phishing emails (Nahmias et al., 2024) and does not contain any real personal data, private communications, or information about real individuals. All annotators are members of the research team and participated voluntarily. The dataset is intended solely for research purposes in intent detection and cybersecurity, and is released under a license that restricts commercial use. We acknowledge that intent detection systems trained on this data could in principle be misused; we encourage responsible use and downstream evaluation on diverse, real-world corpora before deployment.

## 8. Bibliographical References

John Langshaw Austin. 1962. *How to Do Things with Words*. Clarendon Press, Oxford [Eng.].

Davide Bassi, Søren Fomsgaard, and Martín Pereira-Fariña. 2024. Decoding persuasion: a survey on ML and NLP methods for the study of online persuasion. *Frontiers in Communication*, 9.

Paul N. Bennett and Jaime Carbonell. 2005. Detecting action-items in e-mail. In *Proceedings of*

*the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, page 585–586, New York, NY, USA. Association for Computing Machinery.

Caleb T Carr and Rebecca A Hayes. 2017. Social media: Defining, developing, and divining. *Atlantic Journal of Communication*, 23(1):46–65.

Vitor R. Carvalho and William W. Cohen. 2005. On the collective classification of email "speech acts". In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, page 345–352, New York, NY, USA. Association for Computing Machinery.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 38–45, Online. Association for Computational Linguistics.

Robert B. Cialdini. 1993. *Influence: the psychology of persuasion*, rev. edition. Quill, New York.

William W. Cohen, Vitor R. Carvalho, and Tom M. Mitchell. 2004. Learning to classify email into "speech acts". In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 309–316, Barcelona, Spain. Association for Computational Linguistics.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, et al. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. In *Privacy in Machine Learning and Artificial Intelligence Workshop, ICML*.

Giovanni Da San Martino, Alberto Barrón-Cedeño, and Preslav Nakov. 2019. Findings of the NLP4IF-2019 shared task on fine-grained propaganda detection. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 162–170.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. Semeval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1377–1414.

Laura A. Dabbish and Robert E. Kraut. 2006. Email overload at work: an analysis of factors associated with email strain. In *Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work*, CSCW '06, page 431–440, New York, NY, USA. Association for Computing Machinery.

Laura A. Dabbish, Robert E. Kraut, Susan Fussell, and Sara Kiesler. 2005. Understanding email use: predicting action on a message. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '05, page 691–700, New York, NY, USA. Association for Computing Machinery.

Federal Bureau of Investigation. 2023. Internet crime complaint center (ic3) | business e-mail compromise (bec) | ic3.gov.

Ana Ferreira and Soraia Teles. 2019. Persuasion: How phishing emails can influence users and bypass security measures. *Int. J. Hum.-Comput. Stud.*, 125(C):19–31.

Michael C. Frank and Noah D. Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998.

David Gragg. 2003. A multi-level defense against social engineering.

Charles T. Hemphill, John J. Godfrey, and George R. Doddington. 1990. The ATIS spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27,1990*.

Bryan Klimt and Yiming Yang. 2004. The enron corpus: A new dataset for email classification research. In *Machine Learning: ECML 2004*, pages 217–226, Berlin, Heidelberg. Springer Berlin Heidelberg.

Andrew Lampert, Robert Dale, and Cecile Paris. 2010. Detecting emails containing requests for action. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 984–992.

Daniel Nahmias, Gal Engelberg, Dan Klein, and Asaf Shabtai. 2024. Prompted contextual vectors for spear-phishing detection. *ArXiv*, abs/2402.08309.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Jiaxin Pei, Aparna Ananthasubramaniam, Xingyao Wang, Naitian Zhou, Apostolos Dedeloudis, Jackson Sargent, and David Jurgens. 2022. POTATO: The portable text annotation tool. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 327–337, Abu Dhabi, UAE. Association for Computational Linguistics.

Maya Sappelli, Gabriella Pasi, Suzan Verberne, Maaike de Boer, and Wessel Kraaij. 2016. Assessing e-mail intent and tasks in e-mail messages. *Information Sciences*, 358:1–17.

John R. Searle. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press.

Frank Stajano and Paul Wilson. 2011. Understanding scam victims: seven principles for systems security. *Commun. ACM*, 54(3):70–75.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 930–957, Miami, Florida, USA. Association for Computational Linguistics.

Alexandra Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Wei Wang, Saghar Hosseini, Ahmed Hassan Awadallah, Paul N. Bennett, and Chris Quirk. 2019. Context-aware intent identification in email conversations. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 585–594.

Liu Yang, Susan T. Dumais, Paul N. Bennett, and Ahmed Hassan Awadallah. 2017. Characterizing and predicting enterprise email reply behavior. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, page 235–244, New York, NY, USA. Association for Computing Machinery.

## 9. Language Resource References

Oard, Douglas W and Webber, William and Kirsch, David and Golitsynskiy, Sergey. 2015. *Avocado Research Email Collection*. Philadelphia: Linguistic Data Consortium, ISLRN 102-408-869-995-0. PID https://catalog.ldc.upenn.edu/LDC2015T03.

## A. Full Taxonomy Diagram

| Intent Detection Taxonomy | |
|---|---|
| **1. Explicit Intents** | **2. Implicit Intents** |
| • **1.1 Information Manipulation**<br>– 1.1.1 Information Gathering<br>– 1.1.2 Information Control<br>• **1.2 Resource Acquisition**<br>– 1.2.1 Direct Financial Gain<br>– 1.2.2 Indirect Resource Acquisition<br>• **1.3 Trust Establishment**<br>– 1.3.1 Identity Establishment<br>– 1.3.2 Legitimacy Creation<br>• **1.4 Call to Action**<br>– 1.4.1 Direct Action Requests<br>– 1.4.2 Indirect Action Facilitation | • **2.1 Authority**<br>• **2.2 Social Proof**<br>– 2.2.1 Herd Mentality<br>– 2.2.2 Diffusion of Responsibility<br>– 2.2.3 Moral Duty<br>• **2.3 Liking, Similarity, and Deception**<br>– 2.3.1 General Deception<br>– 2.3.2 Deceptive Relationships<br>– 2.3.3 Liking and Similarity<br>• **2.4 Distraction**<br>– 2.4.1 Scarcity (Time / Resource)<br>– 2.4.2 Overloading (Time-based / Information)<br>– 2.4.3 Strong Affect<br>– 2.4.4 Need and Greed<br>• **2.5 Commitment, Integrity, and Reciprocation**<br>– 2.5.1 Integrity<br>– 2.5.2 Consistency<br>– 2.5.3 Commitment (Coercion / Other)<br>– 2.5.4 Reciprocation |

Table 2: Complete hierarchical organization of the intent detection taxonomy.

## B. Sample Analysis Examples

| Communication Excerpt | Explicit Intents | Implicit Intents |
|---|---|---|
| "Your account has been locked due to suspicious activity. Click here to verify your identity immediately to prevent permanent suspension." | 1.1.1 Information Gathering<br>1.4.1 Direct Action Request | 2.1 Authority<br>2.4.1.1 Time Scarcity<br>2.4.3 Strong Affect |
| "As per our previous conversation, I need you to purchase some gift cards for an employee appreciation event. This is urgent and confidential." | 1.2.2 Indirect Resource Acquisition<br>1.4.1 Direct Action Request | 2.1 Authority<br>2.4.1.1 Time Scarcity<br>2.5.2 Consistency |
| "I'm reaching out from accounting regarding your recent invoice. Could you please confirm your banking details for payment processing?" | 1.1.1 Information Gathering<br>1.3.2 Legitimacy Creation | 2.1 Authority<br>2.5.1 Integrity<br>2.5.4 Reciprocation |

Table 3: Sample intent analysis on spear-phishing email excerpts from the DIDECO corpus.

## C. Annotation Guidelines (Excerpt)

The full annotation guidelines are available in the DIDECO repository.[2] Below is a representative excerpt illustrating the structure and format used during the annotation campaign.

**Core Principles.** Each communication may contain multiple intentions, both explicit and implicit. Intentions must be identified based on objective criteria; annotation must remain consistent across annotators; and in cases of ambiguity, annotators are asked to document their reasoning.

**Annotation Process.** For each communication, annotators: (1) read the full message carefully; (2) identify explicit intents (direct communicative goals); (3) identify implicit intents (indirect influence mechanisms); (4) assign each identified intent to its category and subcategory; (5) document justification for ambiguous cases.

**Sample Category Definition — 1.1.1 Information Gathering.** *Definition*: Direct requests aimed at obtaining data, information, or specific details. *Identification criteria*: presence of direct or indirect questions; explicit requests to send information; requests for confirmation or verification. *Example*: *"Can you confirm that the price is the same direct from the factory or from the distributor? If you have the estimates worked up, will you please email them to me as well?"*

**Sample Category Definition — 2.4.3 Strong Affect.** *Definition*: Exploitation of emotions to influence behavior. *Identification criteria*: use of emotionally charged language; appeal to fear, excitement, or curiosity; creation of an emotional state likely to affect judgment. *Example*: *"We have identified a critical security vulnerability in our systems that requires immediate attention. Your swift action will help ensure the safety and integrity of our systems."*

**Complete Annotation Example.** Table 4 illustrates how the annotation format is applied to a phishing email impersonating PayPal.

| **Annotated Text** | **Category** | **Subcategory** |
|---|---|---|
| "We have detected unusual activity on your PayPal account." | 1.3 Trust Establishment | 1.3.2 Legitimacy Creation |
| "Please confirm your information by clicking the link below within the next 24 hours." | 1.4 Call to Action | 1.4.1 Direct Action Request |
| "within the next 24 hours" | 2.4 Distraction | 2.4.1.1 Time Scarcity |
| "If you do not confirm your information, your account will remain limited." | 2.5 Commitment & Reciprocation | 2.5.3.1 Commitment through Coercion |

Table 4: Annotation example on a phishing email excerpt (PayPal impersonation scenario).

[2] https://github.com/Sneusneu/Dideco/

## D. Inter-Annotator Agreement

| Metric | Score |
|---|---|
| Krippendorff's $\alpha$ (overall) | 0.407 |
| Mean pairwise Cohen's $\kappa$ | 0.355 |
| $\kappa$ — Explicit intents | 0.577 |
| $\kappa$ — Implicit intents | 0.111 |

Table 5: Inter-annotator agreement over 169 emails annotated by 2+ annotators. Cohen's $\kappa$ is averaged across all annotator pairs with at least 5 shared emails, binarised at the high-level category level.