Vers une taxonomie pour l'analyse des intentions dans les interactions textuelles numériques

Senaid Popovic[1, 2]
(1) Hornetsecurity, Hem, 59510, France
(2) Université de Lorraine, CNRS, LORIA, F-54000 Nancy, France
senaid.popovic@hornetsecurity.fr

RÉSUMÉ

Cet article propose une taxonomie pour la détection d'intention dans les communications numériques, distinguant les intentions explicites des intentions implicites, basée sur des principes psychologiques de persuasion. Notre approche se distingue par sa capacité à analyser aussi bien les communications numériques légitimes que celles potentiellement malveillantes. Elle repose sur l'identification des intentions sous-jacentes, facilitant ainsi la détection de menaces telles que les arnaques par email (hameçonnage) ou les fraudes sur les réseaux sociaux. Chaque catégorie de la taxonomie est justifiée et illustrée par des exemples de communications correspondant à l'intention associée. Ce travail répond à un manque de ressources dans la recherche sur la détection automatique d'intentions. Il vise à fournir une taxonomie applicable à l'identification des menaces textuelles, notamment les tentatives d'hameçonnage, tout en servant d'outil pédagogique pour sensibiliser le grand public aux stratégies employées dans les communications malveillantes.

ABSTRACT

This article proposes a taxonomy for intent detection in digital communications, distinguishing explicit intentions from implicit intentions, based on psychological principles of persuasion. Our approach is distinguished by its ability to analyze both legitimate digital communications and potentially malicious ones. It relies on identifying underlying intentions, thus facilitating the detection of threats such as email scams or social media fraud. Each category of the taxonomy is justified and illustrated with examples of communications corresponding to the associated intention. This work addresses a lack of resources in research on automatic intent detection. It aims to provide a taxonomy applicable to the identification of textual threats, particularly phishing attempts, while serving as an educational tool to raise public awareness about strategies employed in malicious communications.

MOTS-CLÉS : Détection d'intention, taxonomie, annotation de corpus, ingénierie sociale, llm.

KEYWORDS: intent detection, taxonomy, social engineering, dataset, llm.

ARTICLE : **Soumis à** CORIA-TALN 2025.

1 Introduction

1.1 Contexte et enjeux

L'email est un moyen de communication répandu et essentiel, tant professionnellement que personnellement. En 2021, 319 milliards d'emails étaient envoyés quotidiennement, avec une prévision de croissance continue (The Radicati Group INC., 2021). L'email est particulièrement privilégié dans le cadre professionnel, 86 % des professionnels le considérant comme leur mode de communication préféré (The Radicati Group INC., 2015). Cependant, cette popularité s'accompagne de risques.

Plus de 90 % des cyberattaques proviennent d'emails (Orange Cyberdéfense, 2021), prenant diverses formes comme le spam, hameçonnage (*phishing*), malware ou encore hameçonnage ciblé (*spear-phishing*). Contrairement aux autres types d'attaques, l'hameçonnage ciblé est particulièrement difficile à détecter. D'une part, il est plus rare, ce qui complique la mise en place de systèmes de détection efficaces basés sur des modèles pré-existants. D'autre part, sa surface d'attaque est plus réduite : ces emails ne contiennent souvent ni lien suspect ni pièce jointe malveillante, se limitant à un simple texte soigneusement rédigé pour tromper la victime. Les conséquences financières de ces attaques sont importantes. Selon le rapport du FBI sur la cybercriminalité en 2023 (FBI Internet Crime Complaint Center, 2023), les pertes financières dues aux tentatives de compromission d'email d'entreprise (BEC) ont dépassé $2,9$ milliards de dollars. Ces chiffres soulignent la nécessité de développer des méthodes de détection des intentions malveillantes.

C'est dans ce contexte que notre approche se concentre sur la détection d'intention. Cette approche permet non seulement d'identifier les menaces dissimulées dans les communications textuelles, mais aussi d'offrir des explications aux utilisateurs, les aidant ainsi à mieux comprendre et anticiper ces attaques. En renforçant leur vigilance, elle constitue un outil de sensibilisation et de protection face aux menaces émergentes.

L'email ci-dessous illustre une attaque de type "fraude au président", où l'attaquant usurpe l'identité d'un dirigeant pour inciter la victime à effectuer un virement. Il met en évidence des intentions telles que l'autorité, l'appel à l'action et l'urgence. Bien que nous illustrions ces attaques par email, notre objectif est d'aborder tous les échanges textuels.

> Objet : Virement bancaire
>
> Bonjour Paul, ici le directeur financier.
>
> Je ne suis pas au bureau, mais pourriez-vous effectuer un virement bancaire pour moi aujourd'hui ?
>
> Merci

La notion d'intention constitue un concept complexe à l'intersection de plusieurs disciplines : linguistique, psychologie cognitive, traitement automatique des langues (TAL) et sécurité des systèmes d'information. Contrairement à l'intention telle que définie dans les modèles de compréhension du langage naturel (Natural Language Understanding, NLU) où l'objectif est principalement d'identifier le but opérationnel d'un énoncé, comme pour les assistants vocaux (e.g., Siri, Alexa), l'intention communicative englobe des mécanismes de persuasion, de manipulation et de communication implicite qui dépassent la simple extraction au sens littéral.

1.2 Problématique

Malgré l'importance de cette question, le paysage académique et opérationnel manque d'un cadre pour analyser les intentions communicatives dans leurs différents contextes. Les approches existantes se divisent généralement en deux catégories : les approches sécuritaires, focalisées sur la détection d'hameçonnage et la prévention des attaques d'ingénierie sociale, et les approches organisationnelles, qui étudient les communications entre collaborateurs et la gestion des flux d'information en entreprise.

Ces approches présentent plusieurs limites :

- une analyse textuelle limitée, qui ne prend pas en compte le contexte ni la communication implicite ;
- un manque de ressources annotées pour le développement de systèmes automatiques de détection ;
- une focalisation trop étroite qui néglige la complexité des mécanismes de persuasion.

Un besoin émerge donc : construire une taxonomie fondée sur des principes linguistiques et psychologiques, capable de capturer les intentions communicatives dans des contextes variés, qu'ils soient légitimes ou potentiellement malveillants.

1.3 Objectifs et contributions

Dans ce contexte, notre recherche poursuit trois objectifs :

1. développer une taxonomie des intentions communicatives qui distingue les dimensions explicites et implicites ;
2. élaborer un protocole d'annotation permettant la caractérisation des intentions communicatives ;
3. fournir les outils adaptés à l'utilisation de cette taxonomie dans différents cas, notamment pour la détection de textes malveillants.

Ce papier se focalise sur le premier point et propose une première contribution vers un cadre théorique et méthodologique au travers de cette nouvelle taxonomie.

2 État de l'art

Modèles de classification des intentions communicatives

Les recherches sur les intentions communicatives trouvent leurs racines dans les théories classiques des actes de langage développées par Searle (Searle, 1969). Selon ces travaux, les actes de langage dépassent la simple transmission d'information : ils visent à réaliser des actions et révéler des intentions communicatives sous-jacentes.

Les travaux de Wang et al. (Wang *et al.*, 2019) ont mis en évidence la complexité des échanges professionnels, révélant que les communications s'articulent autour de plusieurs types d'intentions. Les échanges d'informations se structurent entre partage explicite de contenus, demandes de ressources et gestion de tâches collaboratives, chaque interaction portant des intentions communicatives dépassant le simple échange de mots.

Dans le domaine organisationnel, les travaux de Dabbish et al. (Dabbish *et al.*, 2005) ont significativement contribué à la taxonomie des intentions dans les communications professionnelles. Cohen et al. (Cohen *et al.*, 2004) ont proposé des modèles de classification des actes de communication, notamment dans le contexte des emails, mettant en lumière la complexité des intentions communicatives au-delà de leur simple contenu textuel.

Les systèmes de compréhension du langage naturel (NLU) ont développé des approches sophistiquées de détection d'intentions, principalement centrées sur l'extraction de l'objectif communicatif dans les interactions homme-machine (Zhang & Wang, 2016). Ces modèles, bien que performants dans des contextes délimités grâce aux architectures basées sur les transformers et à l'apprentissage profond, peinent à capturer la nuance des intentions implicites, comme le soulignent les recherches récentes (Fan *et al.*, 2024).

Par ailleurs, il convient de mentionner les travaux sur les schémas d'annotation de dialogue comme le schéma DIT++ développé à l'Université de Tilburg par Bunt, qui modélise les intentions dans les interactions dialogiques. Bien que focalisé sur les dialogues oraux, ce schéma offre une perspective complémentaire sur la modélisation des intentions communicatives.

Principes de persuasion et manipulation

Les travaux de Cialdini (Cialdini, 2007) constituent un cadre théorique pour comprendre les mécanismes d'influence. Sa recherche identifie des principes psychologiques clés qui guident les interactions : l'autorité, qui conditionne les individus à répondre à certaines figures ; la rareté, qui valorise ce qui est limité ; la réciprocité, qui crée une obligation de retour après un service ; la preuve sociale, qui incite les individus à suivre ce que font les autres ; la cohérence, qui pousse à maintenir une ligne de conduite conforme aux engagements précédents ; la sympathie, qui augmente l'influence des personnes que l'on apprécie ; et l'unité, qui souligne l'influence des identités partagées.

Stajano et Wilson (Stajano & Wilson, 2011) approfondissent l'analyse des stratégies de manipulation, mettant en lumière les mécanismes psychologiques subtils exploités dans les communications malveillantes. Leurs travaux démontrent comment la pression temporelle, les appels émotionnels et la création de fausses urgences peuvent influencer significativement les comportements individuels.

Ferreira et al. (Ferreira & Teles, 2019) ont intégré ces approches théoriques pour étudier spécifiquement les mécanismes de persuasion dans les emails d'hameçonnage. Leur étude empirique a analysé 194 objets d'emails frauduleux pour identifier les principes de persuasion les plus fréquemment employés. Leurs résultats montrent que les principes d'autorité, d'affect fort (appel aux émotions), d'intégrité et de réciprocité sont prédominants dans ces communications malveillantes. Cette recherche offre un cadre méthodologique pour l'identification des stratégies d'influence dans les communications numériques et confirme l'efficacité d'une approche intégrant plusieurs théories de la persuasion.

Détection des intentions dans les systèmes de dialogue

La compréhension des intentions constitue un enjeu majeur pour les systèmes de dialogue. Comme le soulignent Tur et al. (Tur *et al.*, 2011), l'objectif fondamental est d'identifier automatiquement l'intention de l'utilisateur et d'extraire les arguments associés.

Les approches récentes combinent des techniques avancées d'apprentissage profond (Zhang & Wang, 2016). Les architectures basées sur les transformers, couplées à des méthodes de classification

multi-tâches (Fan *et al.*, 2024), permettent désormais d'intégrer des caractéristiques contextuelles complexes, dépassant les limitations des modèles traditionnels.

Défis et limitations

Malgré les avancées, le domaine de la détection d'intentions communicatives reste confronté à de multiples limitations. La capture des intentions implicites demeure particulièrement complexe, nécessitant une compréhension approfondie des actes de langage et de leur rôle dans l'identification des urgences, comme démontré par (Laurenti *et al.*, 2022). Cette complexité s'accentue lorsqu'on considère les multiples dimensions contextuelles qui influencent l'interprétation des communications.

À notre connaissance, aucune taxonomie existante ne combine intentions implicites et explicites de manière aussi systématique que notre approche. Cette absence représente l'une de nos motivations principales pour ce travail.

Aussi, les ressources annotées restent limitées, notamment pour les langues autres que l'anglais. Ce problème est mis en évidence par (Zhang & Wang, 2016) qui ont dû créer leur propre jeu de données en chinois (CQUD) face à l'absence d'alternatives non-anglophones. Cette limitation freine considérablement le développement de systèmes multilingues robustes et l'extension des recherches à différents contextes culturels.

D'autre part, les systèmes actuels peinent à décoder les nuances des intentions. Les travaux de (Fan *et al.*, 2024) soulignent les difficultés inhérentes aux processus d'annotation et les limites des méthodes existantes qui ne s'adaptent pas aisément à de nouveaux domaines. Par ailleurs, les recherches de (Laurenti *et al.*, 2022) montrent combien cette problématique devient critique dans des contextes d'urgence où la détection précise des intentions est une problématique clé.

3 Taxonomie proposée

3.1 Structure et principes généraux

Notre taxonomie se distingue des approches existantes par deux innovations principales : **(1) la distinction entre intentions explicites et implicites**, et **(2) l'adoption d'une perspective à double usage** reconnaissant qu'un même schéma communicatif peut servir des fins légitimes ou malveillantes selon le contexte.

La taxonomie s'organise autour de la séparation entre intentions explicites et intentions implicites dans les communications numériques. Cette distinction, qui constitue notre contribution principale, permet une analyse plus nuancée des mécanismes communicatifs que les taxonomies binaires (malveillant/non-malveillant) utilisées habituellement.

Notre taxonomie s'appuie sur les travaux de Ferreira (Ferreira & Teles, 2019) sur les principes de persuasion en ingénierie sociale, particulièrement pour la classification des intentions implicites. Nous avons étendu ce cadre pour inclure à la fois les communications malveillantes et légitimes. Le champ d'application se concentre sur les échanges textuels entre deux parties, comme les emails ou messages instantanés, où l'une des parties peut avoir des intentions malveillantes.

Les intentions explicites représentent les objectifs communicatifs directement exprimés qui visent à influencer le comportement du destinataire ou à obtenir des résultats spécifiques. Elles sont classées

en quatre catégories principales : la manipulation d'information, l'acquisition de ressources, l'établissement de confiance, et l'appel à l'action. Ces catégories sont conçues pour couvrir l'ensemble des buts communicatifs directs qu'un émetteur peut poursuivre.

Les intentions implicites, quant à elles, exploitent des principes psychologiques et sociaux pour influencer indirectement le comportement. Regroupées en cinq catégories principales — l'autorité, la preuve sociale, l'affinité/similarité/tromperie, la distraction, et l'engagement/intégrité/réciprocité — elles représentent les mécanismes subtils d'influence qui opèrent au niveau subconscient du destinataire. Des exemples concrets pour chaque catégorie sont fournis en Annexe B.

3.2 Justification des choix taxonomiques

Le choix de nos catégories repose sur trois critères principaux : (1) la fréquence d'observation dans les communications malveillantes documentées dans la littérature, (2) la facilité de détection automatique basée sur des indices linguistiques identifiables, et (3) l'utilité pratique pour la sensibilisation des utilisateurs.

Concernant la fréquence, nos catégories d'intentions explicites correspondent aux objectifs les plus couramment observés dans les attaques documentées : 89% des attaques BEC impliquent une acquisition de ressources, 76% utilisent l'établissement de confiance, et 68% incluent des appels à l'action directs (FBI Internet Crime Complaint Center, 2023). Pour les intentions implicites, l'étude de Ferreira et al. (Ferreira & Teles, 2019) montre que l'autorité (43% des cas), l'affect fort (38%) et l'intégrité (31%) sont les principes les plus fréquemment exploités.

Du point de vue de la détection automatique, chaque catégorie présente des marqueurs linguistiques spécifiques : les demandes d'information contiennent des interrogatives directes, les appels à l'autorité utilisent des titres et des références hiérarchiques, la rareté temporelle s'exprime par des adverbes d'urgence. Cette détectabilité constitue un avantage pratique pour le développement de systèmes automatiques.

Cette granularité permet une sensibilisation ciblée des utilisateurs. Plutôt qu'une simple classification binaire (malveillant/non-malveillant), notre taxonomie explicite les mécanismes d'influence, ce qui facilite la compréhension et la prévention. Par exemple, identifier une "demande d'action sous pression temporelle" est plus informatif qu'une simple alerte "message suspect".

3.3 Intentions explicites

Les intentions explicites dans les communications numériques sont des objectifs clairement exprimés qui visent à influencer le comportement du destinataire ou à obtenir des résultats précis. En nous basant sur la théorie des actes de langage (Searle, 1969), nous proposons quatre catégories qui couvrent les principaux objectifs des communications numériques explicites.

Manipulation de l'information Cette catégorie est présente dans les recherches sur les interactions professionnelles (Dabbish *et al.*, 2005) et dans les études de cybersécurité (Ferreira & Teles, 2019), montrant l'importance de l'échange et du contrôle de l'information dans les communications numériques. La sous-catégorie *Collecte d'informations* inclut les actions visant à obtenir des données, des vérifications légitimes aux tentatives malveillantes. Le *Contrôle de l'information* concerne les

tentatives de gérer le flux d'informations ou l'accès au système, distinction soutenue par les recherches sur l'hameçonnage, qui commence souvent par collecter des informations avant de manipuler le système (Gragg, 2003).

Acquisition de ressources Cette catégorie vient des analyses des transactions commerciales légitimes (Dabbish & Kraut, 2006) et des fraudes (Stajano & Wilson, 2011), reconnaissant que les demandes de ressources sont un objectif distinct de la simple communication d'informations. Nous distinguons le *Gain financier direct*, qui implique des transferts d'argent immédiats, et l'*Acquisition indirecte de ressources*, qui comprend les demandes d'achats ou de services. Cette distinction est confirmée par les recherches sur les fraudes par email professionnel, qui utilisent souvent des demandes financières directes et des manipulations plus subtiles (FBI Internet Crime Complaint Center, 2023).

Établissement de la confiance Cette catégorie mérite sa propre classification d'après des études montrant comment les acteurs légitimes et malveillants tentent d'établir leur crédibilité par des communications directes (Gragg, 2003). La sous-catégorie *Établissement d'identité* concerne les efforts pour vérifier l'identité de l'expéditeur, tandis que la *Création de légitimité* se concentre sur l'établissement du contexte et de la crédibilité via des références à des organisations/entreprises connues (Paypal, Microsoft, etc.). Cette distinction reconnaît que l'identité personnelle et le contexte organisationnel fonctionnent comme des signaux de confiance distincts mais complémentaires.

Appel à l'action En nous appuyant sur les travaux antérieurs de détection d'éléments d'action (Bennett & Carbonell, 2005), nous identifions *l'Appel à l'action* comme représentant un objectif communicatif unique, distinct des stratégies de conformité à long terme. Les *Demandes d'action directes* impliquent des instructions explicites pour des actions immédiates, comme "Cliquez sur ce lien pour réinitialiser votre mot de passe maintenant", tandis que la *Facilitation d'action indirecte* englobe les étapes préparatoires qui permettent ou nécessitent des actions futures, telles que "Conservez ces identifiants pour votre prochaine connexion". Cette distinction est importante pour comprendre la dimension temporelle des intentions de communication numérique, où les actions immédiates sont souvent précédées d'un travail préparatoire qui facilite le comportement ultérieur.

3.4 Intentions implicites

Les intentions implicites exploitent des principes psychologiques et sociaux pour influencer indirectement le comportement. Notre taxonomie adopte la classification établie par Ferreira et al. (Ferreira & Teles, 2019), qui intègre les travaux fondamentaux de Cialdini (Cialdini, 2007), Gragg (Gragg, 2003) et Stajano & Wilson (Stajano & Wilson, 2011) dans une structure unifiée. Cette approche nous permet de bénéficier d'un cadre conceptuel déjà validé pour l'analyse des intentions implicites dans les communications.

Autorité Cette catégorie exploite la tendance des individus à se soumettre aux figures d'autorité ou aux signaux d'autorité. Comme l'ont démontré Ferreira et al. (Ferreira & Teles, 2019), les individus sont conditionnés à répondre positivement aux demandes provenant de sources perçues comme légitimes ou autoritaires.

Preuve sociale Ces intentions utilisent la propension des individus à déterminer leur comportement en observant celui des autres. Ferreira et al. (Ferreira & Teles, 2019) distinguent trois manifestations : la mentalité de troupeau (conformité au groupe), la diffusion de responsabilité (diminution de la responsabilité individuelle en contexte collectif), et le devoir moral (appel aux comportements altruistes).

Affinité, Similarité et Tromperie Cette catégorie regroupe les techniques liées à l'attraction interpersonnelle et aux mécanismes de tromperie. La taxonomie de Ferreira et al. (Ferreira & Teles, 2019) distingue la tromperie générale des relations trompeuses, ainsi que l'exploitation des affinités réelles.

Distraction Ces techniques détournent l'attention de l'évaluation critique. Selon Ferreira et al. (Ferreira & Teles, 2019), elles comprennent la rareté (temporelle ou matérielle), la surcharge cognitive, l'affect fort (exploitation des émotions) et l'appel aux besoins/à l'avidité.

Engagement, Intégrité et Réciprocité Cette catégorie englobe les techniques qui exploitent la cohérence psychologique et les normes d'échange social. Ferreira et al. (Ferreira & Teles, 2019) y incluent l'intégrité (présomption d'honnêteté), la consistance (cohérence avec les actions passées), l'engagement (par coercition ou par d'autres facteurs), et la réciprocité (obligation de rendre une faveur).

4 Discussion et perspectives

4.1 Analyse critique de la taxonomie

La taxonomie proposée constitue une première étape dans l'analyse des intentions communicatives dans les environnements numériques. Néanmoins, plusieurs limites méritent d'être discutées.

Tout d'abord, même si la distinction est faite dans la taxonomie entre intentions explicites et implicites (voir A), dans la réalité, ces notions peuvent être entremêlées et rendre subjective l'annotation des données. Prenons l'exemple d'un message typique :

> "Bonjour, j'espère que vous allez bien. Notre directeur m'a chargé de recueillir les coordonnées personnelles de chaque membre de l'équipe pour mettre à jour l'annuaire d'urgence. Pourriez-vous me transmettre votre adresse personnelle, numéro de téléphone mobile et adresse email privée dès que possible ?"

Ce message peut être interprété comme une simple collecte d'information légitime ou comme une tentative de manipulation utilisant l'autorité et l'intégrité, selon l'expérience et la sensibilité de l'annotateur. Sans contexte supplémentaire, il devient difficile de déterminer objectivement l'intention véritable, illustrant ainsi la complexité de ce type d'analyse.

Cette ambiguïté soulève la question de l'utilité pratique d'une granularité fine par rapport à une approche binaire (malveillant/non-malveillant). Notre choix d'une taxonomie détaillée se justifie

par trois avantages : **(1)** elle permet une sensibilisation ciblée des utilisateurs en explicitant les mécanismes d'influence spécifiques, **(2)** elle facilite le développement de systèmes de détection adaptatifs capables d'identifier des schémas de persuasion émergents, et **(3)** elle offre une base pour des systèmes d'explication permettant aux utilisateurs de comprendre pourquoi un message est potentiellement suspect.

Nous reconnaissons que cette granularité implique un coût d'annotation plus élevé et une complexité accrue pour la classification automatique. Le défi réside dans l'équilibre entre richesse descriptive et praticabilité opérationnelle. Pour adresser cette tension, nous envisageons une approche hiérarchique où les systèmes peuvent opérer à différents niveaux de granularité selon les besoins : détection binaire pour les alertes rapides, et analyse fine pour la formation et la sensibilisation.

Aussi, la dimension temporelle des intentions est laissée de côté pour l'instant. Les communications numériques s'inscrivent souvent dans des séquences d'interactions, où les intentions évoluent. Par exemple, dans une attaque d'hameçonnage ciblé multi-étapes, l'attaquant commence généralement par une phase d'établissement de confiance (première intention), puis passe à un appel à l'action plus direct (seconde intention) une fois que la victime a répondu au message initial. Cette progression séquentielle d'intentions (établissement de confiance $\rightarrow$ appel à l'action) est particulièrement efficace car la réponse de la victime au premier message constitue déjà un indicateur d'engagement qui augmente les chances de succès de l'étape suivante. Notre taxonomie actuelle, bien que statique, pourrait être étendue pour modéliser ces chaînes d'intentions et leur évolution au fil du temps.

4.2 Vers une méthodologie d'annotation systématique

L'étape suivante de notre travail consistera à développer et à appliquer une méthodologie d'annotation pour constituer un corpus de référence selon notre taxonomie. Cette démarche présente plusieurs défis méthodologiques et pratiques que nous envisageons d'aborder de manière systématique.

Notre protocole d'annotation définira les critères d'identification pour chaque catégorie d'intention. Ce protocole comprendra des directives claires et des exemples de référence, permettant une interprétation cohérente des phénomènes observés. La complexité inhérente à l'annotation des intentions implicites, mise en évidence dans notre analyse critique, nécessitera une attention particulière à la formation des annotateurs et à la résolution des cas ambigus.

Nous prévoyons de constituer un corpus initial de 2000 à 3000 emails diversifiés (communications professionnelles légitimes, tentatives d'hameçonnage documentées, emails commerciaux) annotés par une équipe de 3 à 4 annotateurs spécialisés. Cette taille représente un compromis entre faisabilité et représentativité statistique, permettant d'évaluer la robustesse de notre taxonomie tout en restant dans des limites budgétaires raisonnables.

Le défi majeur de cette approche réside dans le coût humain et temporel qu'elle représente. L'annotation manuelle, bien qu'efficace en termes de résultats, devient extrêmement chronophage lorsqu'elle est appliquée à des corpus volumineux. Pour surmonter cette limitation, nous explorerons une approche semi-automatique. En effet, l'annotation assistée par des modèles de langage constitue une piste prometteuse. Nos expérimentations préliminaires avec des LLM comme annotateurs suggèrent qu'ils peuvent identifier efficacement différentes catégories d'intentions, même celles relevant de l'implicite. Cette approche hybride pourrait réduire le temps d'annotation tout en maintenant une qualité acceptable via une validation humaine ciblée.

Concernant la qualité des annotations, nous établirons des métriques d'accord inter-annotateurs adaptées à notre cadre taxonomique. Au-delà du traditionnel coefficient Kappa de Cohen, nos premières réflexions portent sur le développement d'un indice de confiance, qui serait mesuré soit lors de l'annotation manuelle, soit lors de l'annotation via LLM, et qui permettrait d'évaluer la qualité de cette annotation.

La constitution de ce corpus annoté représentera une ressource précieuse pour la communauté scientifique et constituera le fondement des prochaines phases de notre recherche, notamment pour le développement de modèles de détection automatique d'intentions dans les communications numériques.

4.3 Applications potentielles

Les applications pratiques de notre taxonomie s'étendent à plusieurs domaines stratégiques. En cybersécurité, elle offre un cadre analytique pour développer des systèmes de détection d'ingénierie sociale plus avancés, capables d'identifier les schémas de persuasion subtils souvent utilisés dans les attaques ciblées. L'approche multi-niveaux de notre taxonomie permet de dépasser les limitations des systèmes actuels, principalement basés sur des analyses lexicales et syntaxiques (Ferreira & Teles, 2019).

Notre taxonomie peut aussi servir de base à des systèmes d'explication, alertant les utilisateurs sur les potentielles tentatives de manipulation ou aidant à clarifier les intentions communicatives dans les échanges textuels. Ainsi, cette taxonomie pourrait constituer un outil de sensibilisation et de formation à la communication numérique sécurisée, permettant aux utilisateurs d'identifier et de comprendre les stratégies de persuasion potentiellement malveillantes.

Au-delà de la cybersécurité, notre approche trouve des applications en analyse des réseaux sociaux, où l'identification des intentions communicatives peut aider à détecter la désinformation ou les campagnes d'influence. Elle peut également contribuer à l'amélioration des systèmes de dialogue en permettant une meilleure compréhension des intentions utilisateur dans leurs dimensions explicites et implicites.

4.4 Prochaines étapes du projet

Notre programme de recherche s'articule autour de deux axes prioritaires pour les développements futurs.

Le premier axe concerne l'enrichissement de notre corpus annoté. Nous prévoyons d'étendre nos annotations à un ensemble plus diversifié de communications numériques, incluant des emails professionnels légitimes, des tentatives d'hameçonnage sophistiquées et des communications issues de réseaux sociaux. Cette diversification permettra de tester la robustesse de notre taxonomie dans différents contextes communicationnels. L'objectif est de constituer un corpus de plusieurs milliers de communications annotées semi-automatiquement par des grands modèles de langage.

Le deuxième axe vise le développement de modèles automatiques de détection d'intention. En nous appuyant sur notre corpus annoté, nous explorerons l'efficacité de différentes architectures d'apprentissage profond. Une attention particulière sera portée à l'interprétabilité de ces modèles, mais aussi à leur frugalité, permettant leur déploiement dans des environnements contraints.

5 Conclusion

Dans cet article, nous présentons une taxonomie pour l'analyse et la détection des intentions communicatives dans les interactions textuelles numériques, comblant ainsi une lacune dans la littérature scientifique à l'intersection de la linguistique, de la psychologie sociale et de la cybersécurité. Notre contribution principale réside dans la proposition d'un cadre qui distingue systématiquement les intentions explicites et implicites, tout en reconnaissant leur complémentarité et leur utilisation potentielle dans des contextes légitimes comme malveillants.

Les fondements théoriques de notre taxonomie, ancrés dans les théories des actes de langage et les principes psychologiques de persuasion, lui confèrent une base conceptuelle solide. L'articulation des catégories d'intentions, notamment au niveau implicite, permet de capturer la complexité des mécanismes persuasifs à l'œuvre dans les communications. La structure complète de cette taxonomie est présentée en annexe (voir A).

Le développement d'un protocole d'annotation constituera une première étape vers la création d'un corpus de référence qui pourra servir de base au développement de systèmes automatiques de détection d'intention. Les défis méthodologiques identifiés, notamment concernant l'annotation à grande échelle et la subjectivité inhérente à l'interprétation des intentions implicites, ouvrent de nouvelles perspectives de recherche.

Les applications de nos travaux dépassent le cadre de la cybersécurité pour s'étendre aux systèmes de dialogue, à l'analyse des réseaux sociaux et à la formation des utilisateurs. Cette taxonomie répond à un manque de ressources dans la recherche autour de la détection d'intention, offrant un cadre conceptuel et méthodologique qui pourra bénéficier tant aux chercheurs qu'aux développeurs de solutions de sécurité numérique.

Les travaux futurs porteront sur la validation de cette taxonomie par l'annotation d'un corpus représentatif, le développement de modèles de détection automatique, et l'évaluation de son efficacité dans des contextes applicatifs réels.

Références

BENNETT P. & CARBONELL J. (2005). Detecting action-items in e-mail. p. 585–586. DOI : 10.1145/1076034.1076140.

CIALDINI R. B. (2007). *Influence : The Psychology of Persuasion*. Harper Business, revision edition édition.

COHEN W., CARVALHO V. & MITCHELL T. (2004). Learning to classify email into "speech acts". In *Proceedings of Empirical Methods in Natural Language Processing*.

DABBISH L. A. & KRAUT R. E. (2006). Email overload at work : An analysis of factors associated with email strain. In *CSCW '06*.

DABBISH L. A., KRAUT R. E., FUSSELL S. & KIESLER S. (2005). Understanding email use : Predicting action on a message. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* : ACM.

FAN L., PU J., ZHANG R. & WU X.-M. (2024). Lanid : Llm-assisted new intent discovery. In *LREC-COLING 2024*.

FBI INTERNET CRIME COMPLAINT CENTER (2023). *Internet Crime Report 2023*. Rapport interne, FBI.

FERREIRA A. & TELES S. (2019). Persuasion : how phishing emails can influence users and bypass security measures. *International Journal of Human-Computer Studies*, **125**, 19–31.

GRAGG D. (2003). *A Multi-Level Defense Against Social Engineering*. Rapport interne, SANS Institute - InfoSec Reading Room.

LAURENTI E., BOURGON N., BENAMARA F., MARI A., MORICEAU V. & COURGEON C. (2022). Speech acts and communicative intentions for urgency detection. In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics* : Association for Computational Linguistics.

ORANGE CYBERDÉFENSE (2021). Article "phishing et gestion des emails, nos conseils".

SEARLE J. R. (1969). *Speech Acts : An Essay in the Philosophy of Language*. Cambridge University Press.

STAJANO F. & WILSON P. (2011). Understanding scam victims : Seven principles for systems security. *Communications of the ACM*, **54**(3), 70–75.

THE RADICATI GROUP INC. (2015). Email statistics report.

THE RADICATI GROUP INC. (2021). Email statistics report.

TUR G., HAKKANI-TÜR D., HECK L. & PARTHASARATHY S. (2011). Sentence simplification for spoken language understanding. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* : IEEE.

WANG B., LIAKATA M., ZUBIAGA A. & PROCTER R. (2019). Mining user intents in twitter : A semi-supervised approach to inferring intent categories for tweets. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics* : ACL.

ZHANG X. & WANG H. (2016). A joint model of intent determination and slot filling for spoken language understanding. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*.

A Taxonomie complète

La taxonomie présentée dans cet article est résumée dans son intégralité ci-dessous.

Intentions explicites	Intentions implicites
1.1 Manipulation d'information	**2.1 Autorité**
1.1.1 Collecte d'information	
1.1.2 Contrôle d'information	
1.2 Acquisition de ressources	**2.2 Preuve sociale**
1.2.1 Gain financier direct	2.2.1 Mentalité de troupeau
1.2.2 Acquisition indirecte de ressources	2.2.2 Diffusion de responsabilité
	2.2.3 Devoir moral
1.3 Établissement de confiance	**2.3 Affinité, similarité et tromperie**
1.3.1 Établissement d'identité	2.3.1 Tromperie générale
1.3.2 Création de légitimité	2.3.2 Relations trompeuses
	2.3.3 Affinité et similarité
1.4 Appel à l'action	**2.4 Distraction**
1.4.1 Demandes d'action directes	2.4.1 Rareté
1.4.2 Facilitation d'action indirecte	2.4.1.1 Rareté temporelle
	2.4.1.2 Rareté de ressources
	2.4.2 Surcharge
	2.4.2.1 Surcharge temporelle
	2.4.2.2 Surcharge informationnelle
	2.4.3 Affect fort
	2.4.4 Besoin et avidité
	2.5 Engagement, intégrité et réciprocité
	2.5.1 Intégrité
	2.5.2 Cohérence
	2.5.3 Engagement
	2.5.3.1 Engagement par coercition
	2.5.3.2 Engagement par d'autres facteurs
	2.5.4 Réciprocité

TABLE 1 – Vue d'ensemble de la taxonomie des intentions communicatives

B Exemples détaillés par catégorie de la taxonomie

Cette section présente des exemples concrets pour chaque catégorie de la taxonomie, illustrant comment les intentions explicites et implicites se manifestent dans les communications numériques.

B.1 Intentions explicites

Catégorie	Exemple
1.1 Manipulation d'information	
1.1.1 Collecte d'information	"Pourriez-vous m'envoyer le document de spécifications pour le nouveau projet ?"
1.1.2 Contrôle d'information	"Veuillez mettre à jour vos paramètres de sécurité selon les nouvelles directives."
1.2 Acquisition de ressources	
1.2.1 Gain financier direct	"Veuillez transférer 500€ au compte bancaire suivant pour finaliser votre commande."
1.2.2 Acquisition indirecte de ressources	"Merci d'acheter des cartes-cadeaux pour l'événement d'appréciation des employés."
1.3 Établissement de confiance	
1.3.1 Établissement d'identité	"Je suis le nouveau responsable informatique pour votre région et je superviserai la mise à jour des systèmes."
1.3.2 Création de légitimité	"Notre société collabore avec les plus grandes entreprises du secteur depuis plus de 15 ans."
1.4 Appel à l'action	
1.4.1 Demandes d'action directes	"Veuillez remplir le formulaire ci-joint avant la fin de la journée."
1.4.2 Facilitation d'action indirecte	"Nous préparons actuellement l'accès au nouveau système et vous enverrons vos identifiants prochainement."

TABLE 2 – Exemples d'intentions explicites dans les communications numériques

B.2 Intentions implicites

Catégorie	Exemple
2.1 Autorité	
2.1 Autorité	"En tant que membre du comité exécutif, j'ai déterminé que tous les départements doivent se conformer aux nouveaux protocoles de sécurité."
2.2 Preuve sociale	
2.2.1 Mentalité de troupeau	"Tous les membres de l'équipe ont déjà complété leur formation sur le nouveau système."
2.2.2 Diffusion de responsabilité	"Votre équipe entière doit participer à cette initiative pour assurer son succès."
2.2.3 Devoir moral	"Votre participation aidera les collègues qui sont en difficulté avec le nouveau système."
2.3 Affinité, similarité et tromperie	
2.3.1 Tromperie générale	"La promotion que vous avez gagnée est garantie par notre partenaire de confiance international."
2.3.2 Relations trompeuses	"En tant que développeurs comme vous, nous comprenons parfaitement les défis auxquels vous êtes confrontés."
2.3.3 Affinité et similarité	"J'ai remarqué que nous avons tous deux étudié à l'Université de Lorraine. J'aurais besoin de votre expertise sur ce projet."
2.4 Distraction	
2.4.1 Rareté	"Cette opportunité de formation n'est disponible que pendant les 24 prochaines heures. Seuls trois postes restent disponibles."
2.4.2 Surcharge	"Veuillez consulter les 25 pages du nouveau protocole et répondre d'urgence à cette demande complexe avant la fin de journée."
2.4.3 Affect fort	"ALERTE CRITIQUE : Changements importants dans la politique de sécurité nécessitant votre attention immédiate !"
2.4.4 Besoin et avidité	"Félicitations ! Vous avez été sélectionné pour recevoir un bonus de 5000€. Pour confirmer votre éligibilité, veuillez compléter ce formulaire."
2.5 Engagement, intégrité et réciprocité	
2.5.1 Intégrité	"Je tiens à être totalement transparent avec vous concernant cette situation délicate. J'ai besoin de votre aide."
2.5.2 Cohérence	"En tant que personne ayant toujours valorisé l'excellence dans les projets précédents, votre participation à cette initiative est essentielle."
2.5.3 Engagement	"Conformément à l'accord que vous avez signé la semaine dernière, nous avons besoin que vous complétiez cette étape supplémentaire."
2.5.4 Réciprocité	"Suite au soutien exceptionnel que nous avons fourni à votre équipe le mois dernier, nous espérons pouvoir compter sur votre aide pour cette question urgente."

TABLE 3 – Exemples d'intentions implicites dans les communications numériques