Université de Reims Champagne-Ardenne

École doctorale : "Sciences du Numérique et de l'Ingénieur"

Thèse

pour obtenir le grade de

Docteur de l'Université de Reims Champagne-Ardenne

Discipline : Informatique

Spécialité : Agriculture numérique de précision, Traitement Automatique du Langage Naturel, Apprentissage Machine et Intelligence Artificielle

Présentée et soutenue publiquement par

Shufan JIANG

le mercredi 14 décembre 2022

Intégration de données textuelles pour la détection de risques naturels en agriculture

---

Jury :

Mathieu ROCHE	CIRAD	Rapporteur
Antoine DOUCET	Université de La Rochelle	Rapporteur
Myriam LAMOLLE	Université Paris 8	Présidente
Gülgün KAYAKUTLU	İstanbul Technical University	Examinatrice
Raja CHIKY	ISEP Paris & 3DS OUTSCALE	Co-directrice de thèse
Francis ROUSSEAUX	Université de Reims Champagne-Ardenne	Co-directeur de thèse
Rafael ANGARITA	Université Paris Nanterre	Co-encadrant de thèse
Stéphane CORMIER	Université de Reims Champagne-Ardenne	Co-encadrant de thèse
Julien ORENSANZ	CAP2020	Membre invité

Contents

Remerciements

Ma thèse de doctorat est arrivée à son terme dans les délais grâce à beaucoup de gens. Malheureusement, je ne peux pas citer le nom de chaque personne à qui je suis reconnaissante.

J'aimerais remercier en premier lieu ma co-directrice de thèse, Raja Chiky, qui m'a aidé à trouver cette thèse quand j'avais du mal à entrer dans monde de la recherche, et qui m'a beaucoup soutenue pendant ces trois dernières années.

Je remercie également mon co-directeur de thèse, Francis Rousseaux, pour m'avoir poussée en dehors de ma zone de confort et avoir mené mes pensées vers de plus hauts niveaux d'abstraction.

Je remercie aussi mes co-encadrants – je préfère même l'appellation en allemand "meine doktorväter" (mes pères docteurs) – Rafael Angarita et Stéphane Cormier. En sortant de mon poste d'ingénieuse en 2019, j'avais zéro publication scientifique. C'est avec une immense patience et beaucoup de bonne humeur que vous avez corrigé ma rédaction et aiguisé mes idées. Je vous remercie pour vos disponibilités à m'avoir aidé dans les différentes démarches, malgré vos emplois du temps déjà chargés. Grâce à vos efforts, j'ai pu conduire ma recherche sans être trop perturbée par la pandémie de Covid-19. Les réunions hebdomadaires avec vous étaient toujours agréables, car vous avez su tolérer mes retards et vous m'avez fait confiance. La liberté que vous m'avez accordée a préservé ma curiosité scientifique.

Je remercie particulièrement notre partenaire et expert en agriculture, Julien Orensanz, pour nous avoir identifié l'idée très pertinente de l'intégration des données hétérogènes pour détecter les signaux faibles en agriculture. Pendant trois ans, tu nous as donné des conseils dans le domaine agricole et contribué à l'annotation des données.

Je remercie aussi nos collaborateurs d'Arvalis, qui ont passé du temps avec nous pour trouver des cas d'utilisations, pré-analyser les tweets collectés et labelliser des tweets.

Merci aux rapporteurs Mathieu Roche et Antoine Doucet pour votre lecture attentive et vos conseils très pointus. Merci aux examinatrices Myriam Lamolle et Gülgün Kayakutlu pour votre participation à la soutenance. Elle fut mon meilleur

souvenir de l'année 2022 grâce à la bienveillance et au soutien du jury.

Je remercie tous les enseignants-chercheurs, les élèves qui ont assisté à mes cours et les reviewers anonymes pour m'avoir donné des suggestions ou posé des questions sur ce projet de thèse.

Merci à Lionel Trojman, Céline Delagneau, Ida Lenclume, Marielle Tur et Ségolène Buffet pour m'avoir aidé dans les tâches administratives.

Merci aux membres du service informatique de l'ISEP, notamment Elliot Andrzejewski, Abdnasser Agnaou, Christophe Boyanique et Madjid Habi : vous avez assuré le bon fonctionnement de l'infrastructure nécessaire à mes expériences.

Merci à mes amis doctorants et stagiaires des laboratoires, Maurras Togbe, Jade Guisiano, Mariam Barry, William Aboucaya, Nan Ding, Ekaterina Kalinicheva, Arthur Vervaet, Guillaume Lachaud, Abir Aissa, Clément Royer, Nataly Pozo, Lorenzo Guercio, Guillaume Dupont, Matthieu Pombet, Sahar Husseini, Sung Hyuk Pang, Mohamed El Fatri, Juliet Moso, Mohamed Zohir Koufi et Chaima Messaoudi : c'est toujours un plaisir de discuter avec vous.

Je remercie encore Gérald Pineau et Bruno Houdouin pour m'avoir embauché en 2015 sur le projet du chatbot 2sTalk. Cet agent conversationnel m'a, non seulement donné mon premier emploi en France et au sein d'une équipe adorable, mais aussi cultivé mon intérêt sur la gestion des connaissances, le traitement automatique de langage et l'intelligence artificielle. J'ai été très heureuse de vous voir connectés à la soutenance de cette thèse !

Enfin, un grand merci à mes chères amies Qiu, Saileng, Laoka, 15, 17, 003, DML, NSW, Marie-Carmen et à mon cher compagnon Mathieu pour avoir écouté et lu mes longues palabres. Votre compréhension et soutien m'ont sauvée de mes doutes et inquiétudes.

Résumé étendu en français

Chapitre 1. Introduction

Les progrès récents des technologies de l'information et de la communication (TIC) visent à relever certains des défis les plus importants auxquels l'agriculture est confrontée aujourd'hui [35]. Les chercheurs ont appliqué un large éventail de technologies pour atteindre certains objectifs spécifiques. Parmi ces objectifs : la prévision climatique en agriculture à l'aide de modèles de simulation [63], rendre la production de certains types de céréales plus efficace et efficiente avec la vision par ordinateur et l'intelligence artificielle [124], l'évaluation des sols avec des drones [171], et le paradigme de IoT lorsque des dispositifs connectés tels que des capteurs récolent des données en temps réel au niveau du champ et qui, combinées à l'informatique en nuage, peuvent être utilisées pour surveiller les composants de système de production agricole tels que le sol, les plantes, les animaux, les conditions météorologiques et autres conditions environnementales [123]. L'utilisation de ces TIC pour améliorer les processus agricoles est connue sous le nom de *smart farming* [194].

Face au défi que représentent l'augmentation de la population et l'évolution des habitudes alimentaires, l'agriculture de précision apparaît pour accroître la durabilité de la production alimentaire. En effet, la durabilité de la production alimentaire fait partie de l'objectif « Zéro faim » de l'Agenda 2030 pour le développement durable des Nations Unies [96]. Les questions phytosanitaires, notamment (a) les stress biotiques tels que les mauvaises herbes, les insectes ravageurs, les animaux ou les agents pathogènes nuisibles aux plantes ou aux produits végétaux, et (b) les stress abiotiques tels que les inondations, la sécheresse, les températures extrêmes, peuvent entraîner une perte de production alimentaire. Un sujet essentiel de l'agriculture de précision est l'amélioration des tâches de prévention des risques et la mesure des risques naturels dans leurs aspects globaux et locaux par une surveillance en temps réel.

D'autre part, les agriculteurs pratiquent les technologies de « l'agriculture moderne » depuis la troisième révolution agricole à la fin des années 1960, qui implique des engrais chimiques, des variétés à haut rendement, la mécanisation et l'irriga-

tion. Réduire l'utilisation et l'impact des pesticides pour soutenir le développement durable de l'agriculture de production pourrait réexposer les agriculteurs à l'incertitude du rendement, aux facteurs non contrôlables de la production et à des phénomènes complexes dont ils n'ont pas une connaissance stabilisée [133]. Les approches holistiques traditionnelles peuvent être utiles. De telles approches, toujours guidées par les connaissances locales et l'hybridation des preuves empiriques, sont ignorées pour leur « inefficacité » à court terme. L'utilisation de nouvelles technologies peut introduire de nouveaux verrous socio-technologiques [74], renforçant les procédures de décision dominantes. L'évolution technologique et les changements climatiques exigent une transition des connaissances en agriculture.

Vers l'intégration des données textuelles

Une enquête sur les obstacles à l'application du Big Data dans l'agriculture [191] mentionne les erreurs dans les données, l'inaccessibilité liée au volume des données et au manque de bande passante de communication dans les zones rurales, l'incompatibilité entre les différents entrepôts de données et outils de traitement, et l'inutilisabilité en raison de l'hétérogénéité des données. En effet, dans le contexte de l'agriculture intelligente, les **dispositifs IoT** eux-mêmes sont à la fois producteurs et consommateurs de données et ils produisent des *données hautement structurées*. Les informations importantes liées à l'agriculture peuvent également provenir de différentes sources telles que les journaux et les rapports périodiques officiels comme les Bulletins de Santé des Végétaux (BSV, pour son nom en français *Bulletin de Santé du Végétal*)[1], les médias sociaux comme Twitter et les expériences des agriculteurs.

L'objectif du BSV est de : i) présenter un rapport sur la santé des cultures, incluant leurs stades de développement, les observations de ravageurs et de maladies, et la présence de symptômes qui leur sont liés ; et ii) fournir une évaluation du risque phytosanitaire, en fonction des périodes de sensibilité des cultures et des seuils de ravageurs et de maladies. Le BSV et les autres rapports officiels sont des *données semi-structurées*. La production de connaissances du BSV est la suivante :

1. Le réseau d'observation Epiphyt [2] collecte et centralise les observations agronomiques consécutives réalisées sur l'ensemble du territoire français par les réseaux de surveillance régionaux impliquant 400 observateurs dans 36 réseaux partenaires pour constituer une base de données nationale.

2. Les experts construisent des modèles statistiques [111] pour analyser les données collectées, extraire les informations critiques et faire des prédictions.

[1] https ://agriculture.gouv.fr/bulletins-de-sante-du-vegetal
[2] https ://agroedieurope.fr/wp-content/uploads/fiche-projet-epiphyt-fr.pdf

3. Les processus de génération de rapports compilent les informations et les prédictions en alertes, en langage naturel ou en graphiques pour constituer le contenu du BSV. Au cours de cette étape, les connaissances interprétables par la machine sont transformées en données compréhensibles par l'homme.

Ainsi, les BSV contiennent des connaissances agricoles précieuses en France sur plusieurs décennies.

Twitter - ou tout autre média social - peut être utilisé comme une plateforme d'échange de connaissances sur la gestion durable des sols [113], et il peut également aider le public à comprendre les questions agricoles et soutenir la communication des crises dans l'agriculture [3]. Nous pouvons aussi acquérir **les expériences des agriculteurs** (aussi connue sous le nom d'anciennes pratiques agricoles ou connaissances ancestrales) par les entretiens et les processus participatifs. Les messages sur les médias sociaux et les expériences des agriculteurs sont des *données non structurées*. Nous exprimerons plus en détail la valeur des médias sociaux dans la section suivante.

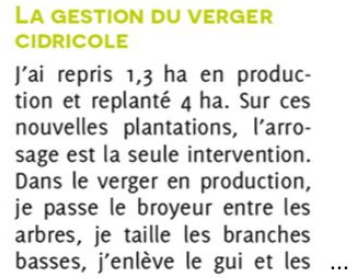
La figure 1 illustre comment ces données hétérogènes provenant de différentes sources peuvent se présenter aux agriculteurs : les informations ne sont pas toujours explicites ou opportunes. Le traitement du langage naturel (NLP) et les graphes de connaissances sont des technologies permettant l'intégration de données, l'extraction d'informations et la reconstruction de connaissances. Nous passons en revue le NLP et les graphes de connaissances dans l'agriculture dans le chapitre 2.

Surveillance de la santé des plantes sur les médias sociaux

Gao et al. [55] classent les technologies existantes de surveillance en temps réel des risques naturels en deux catégories : (i) la surveillance indirecte par l'analyse des paramètres environnementaux produits par les réseaux de capteurs et les dispositifs de l'Internet des objets (IoT) pour déduire la probabilité des risques phytosanitaires [119] ; et (ii), la surveillance directe par le traitement des images [134]. Cependant, les technologies actuelles d'agriculture de précision favorisent les pratiques de monoculture à grande échelle qui sont non durables et économiquement risquées pour les agriculteurs [64]. De plus, selon l'Organisation des Nations unies pour l'alimentation et l'agriculture, les exploitations de moins de 2 hectares représentaient 84% de toutes les exploitations agricoles dans le monde en 2019, et la plupart de ces petites exploitations sont des exploitations familiales [101]. Ainsi, nous suggérons que les données d'observation actuelles de l'agriculture de précision ne peuvent pas représenter toutes les formes d'exploitations agricoles, en particulier les petites exploitations. Récemment, la façon d'encourager la participation des agriculteurs à partager leurs connaissances et leurs observations attire l'attention des chercheurs [84, 89]. Cependant, les observations locales des agriculteurs ne sont

Données non structurées provenant de Twitter

Données semi-structurées : Bulletin de Santé du Végétal



Données non structurées provenant des expériences des agriculteurs

heure_utc	heure_de_paris	temperature	humidite	pression
2019-08-31T18:00:00+00:00	August 31, 2019 8:00 PM	29.8 °C	55 %	99,600 Pa
2019-08-31T16:15:00+00:00	August 31, 2019 6:15 PM	32.7 °C	48 %	99,500 Pa
2019-08-31T11:30:00+00:00	August 31, 2019 1:30 PM	27.3 °C	66 %	99,800 Pa
2019-08-31T09:45:00+00:00	August 31, 2019 11:45 AM	25.2 °C	73 %	99,900 Pa
2019-08-31T15:45:00+00:00	August 31, 2019 5:45 PM	32.8 °C	47 %	99,500 Pa
2019-08-27T04:30:00+00:00	August 27, 2019 6:30 AM	22.1 °C	84 %	99,700 Pa
2019-08-27T05:45:00+00:00	August 27, 2019 7:45 AM	22.2 °C	84 %	99,800 Pa
2019-09-02T11:15:00+00:00	September 2, 2019 1:15 PM	22.2 °C	52 %	100,500 Pa
2019-09-02T16:30:00+00:00	September 2, 2019 6:30 PM	25.4 °C	36 %	100,300 Pa
2019-09-03T01:45:00+00:00	September 3, 2019 3:45 AM	14.5 °C	84 %	100,600 Pa
2019-09-02T18:30:00+00:00	September 2, 2019 8:30 PM	23.6 °C	39 %	100,400 Pa

Données structurées provenant d'une Station météorologique

FIGURE 1 : Sources hétérogènes de données agricoles : *données non structurées* provenant de Twitter et d'expériences d'agriculteurs `www.bio-centre.org` ; *données semi-structurées* provenant des Bulletins de santé des végétaux français ; et *données structurées* provenant d'une station météorologique `www.data.gouv.fr`.

pas prises en compte, ce qui entraîne une perte de légitimité et la disparition des connaissances traditionnelles locales. Comme Ingram [71] le souligne, les connaissances locales des agriculteurs reposent sur des processus sociaux d'échange de connaissances. Or, la réduction du nombre d'agriculteurs et l'individualisme affaiblissent les liens locaux de sang et de voisinage pour l'acquisition des connaissances. La diversification des professions dans le domaine agricole déstabilise également les structures traditionnelles de sociabilité professionnelle [167].

Les médias sociaux comme Twitter jouent un rôle croissant dans l'échange de connaissances entre agriculteurs et entre agriculteurs et professionnels du monde rural. Il semble que l'utilisation de Twitter parmi les professionnels ruraux et les agriculteurs ait bien évoluée, avec une participation ouverte, une collaboration (retweet) et un engagement plus complet (poser des questions, fournir des réponses) dominant la messagerie à sens unique (nouveaux tweets/originaux) [131]. Selon le paradigme de la *détection sociale* [186], les individus - qu'ils soient agriculteurs ou

non - ont de plus en plus de connectivité à l'information lorsqu'ils se déplacent sur le terrain. Chaque individu peut devenir un diffuseur d'informations. En ce sens, les informations sur les risques en temps réel sont publiées sur les réseaux sociaux tels que Twitter. En effet, Twitter permet aux agriculteurs d'échanger leurs expériences, de s'abonner à des sujets d'intérêt à l'aide de hashtags et de partager des informations en temps réel sur les risques naturels. Par rapport aux applications payantes, les informations sur Twitter, présentées sous forme de texte, d'image, de son, de vidéo ou d'un mélange des trois, sont plus accessibles au public, mais moins formalisées ou structurées. De plus en plus d'agriculteurs participent à des communautés Twitter en ligne en ajoutant des hashtags tels que #AgriChatUK (http ://www.agrichatuk.org) ou #FrAgTw (https ://franceagritwittos.com), à leurs messages sur Twitter [41]. Ainsi, nous pouvons considérer Twitter comme un outil ouvert pour l'échange de connaissances entre agriculteurs.

Contributions principales

La hiérarchie données-information-connaissance (DIC) de [150] définit les données, les informations, les connaissances et leurs processus de transformation comme suit :

- Les données sont des valeurs observées d'objets, d'événements et de leur environnement. Les données peuvent être traitées par des ordinateurs.
- L'information est contextualisée, et fournit des réponses aux questions qui, quoi, où et quand. Les informations peuvent être déduites des données.
- La connaissance est un savoir-faire, représentant la capacité à comprendre les informations, puis à les transformer en jugements, opinions, prédictions et décisions.

En suivant la hiérarchie DIC, nous examinons le flux actuel de gestion des connaissances avec les sources de données existantes dans la Figure 2. À ce stade, les données structurées sont centralisées et traitées par des modèles statistiques pour produire des prédictions. Ces informations sont ensuite utilisées pour générer le contenu textuel du BSV. Les agriculteurs peuvent soit fournir des données brutes, soit consommer l'information traitée dans le BSV. Pour les ordinateurs, les connaissances traitables sont les modèles statistiques et les ontologies de domaine [148] qui indexent la date, la région et une partie des cultures mentionnées dans le BSV. Le développement manuel de l'ontologie de domaine limite l'indexation du BSV. En parallèle, bien que certains tweets mentionnent le contenu de certaines BSV, les échanges de connaissances des agriculteurs sur les médias sociaux restent non transmis à la base de connaissances formelle.

La figure 3 illustre la contribution de cette thèse. Nous proposons :

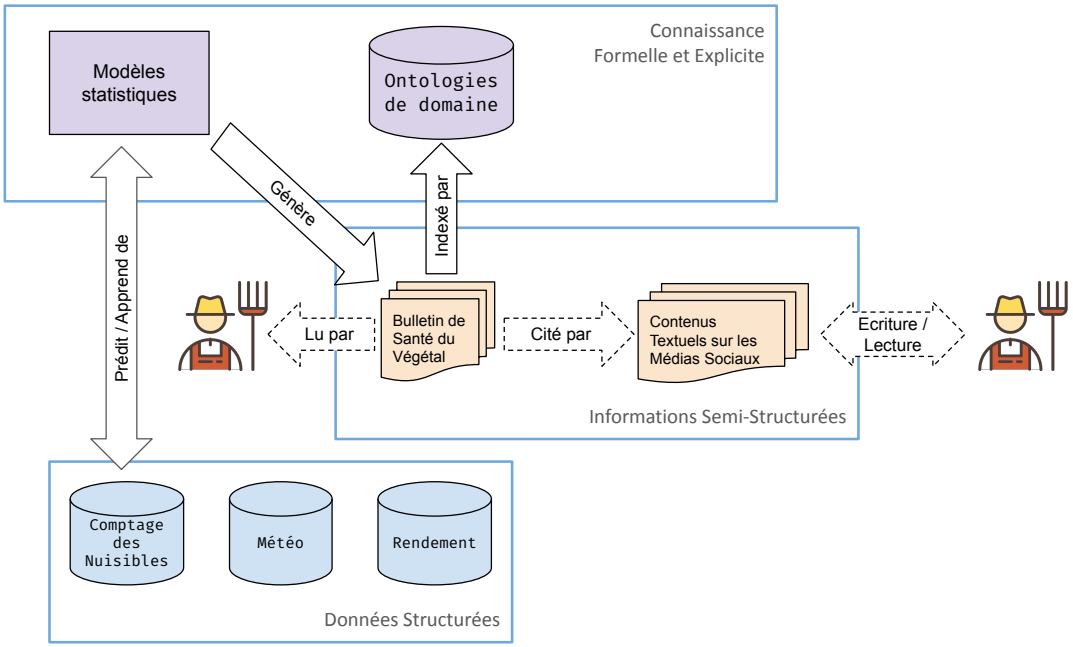


FIGURE 2 : Flux de gestion des connaissances au point de départ.

- Utiliser Twitter comme plateforme ouverte de surveillance de la santé des plantes pour impliquer les agriculteurs dans l'acquisition de connaissances agricoles.
- Un modèle de langage pré-entraîné ChouBERT comme base de connaissances formelle, mais implicite. Cette base de connaissances peut « apprendre automatiquement » des BSV et aider à extraire des informations d'autres données textuelles comme les tweets.

Notre idée est d'impliquer plus de connaissances individuelles et de renforcer la communication entre les praticiens du domaine pour réunir les organisations de recherche, les instituts techniques, les entreprises privées et les agriculteurs et améliorer la vulgarisation. Cette thèse aborde cette lacune en proposant d'utiliser des technologies de fouille de texte pour extraire des informations des médias sociaux et renforcer l'acquisition de connaissances sur la santé des plantes.

Chapitre 2. État de l'art

Le traitement automatique du langage naturel (TALN), ou linguistique informatique, est un sous-domaine de l'informatique, de la linguistique et de l'intelligence

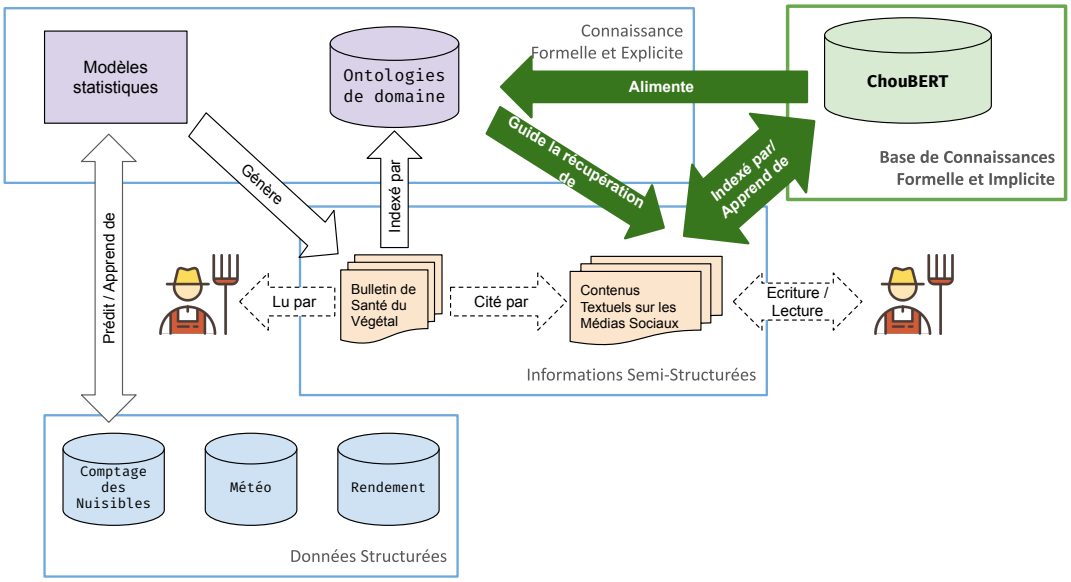


FIGURE 3 : Contribution de Vivace au flux de gestion des connaissances.

artificielle qui étudie comment les ordinateurs peuvent être utilisés pour comprendre, manipuler et produire du contenu en langage humain [33, 66]. Les applications réelles du TALN comprennent la détection du pourriel, la traduction automatique, l'exploration des médias sociaux pour le suivi des questions environnementales ou sociales, l'extraction d'informations des dossiers médicaux électroniques, etc. Dans le contexte de la construction d'une base de connaissances pour la surveillance de la santé des plantes, cette thèse considère deux tâches de TALN :

- Classification au niveau du texte, telle que la distinction entre les observations sur les risques naturels et les autres informations, ou la prédiction du risque basé sur une description textuelle donnée. Selon [44], la classification de textes est « un processus de recherche d'information (RI) qui regroupe les documents dans des classes prédéfinies ». Le processus de RI permet de filtrer les données non pertinentes dans les grandes collections de documents.

- La classification au niveau des jetons, qui vise à classer un ou plusieurs mots dans des classes prédéfinies. Dans notre cas, nous nous intéressons aux agents pathogènes, aux caractéristiques des plantes, aux stades de développement des insectes nuisibles, aux noms des espèces concernées et aux lieux. La classification des entités intéressées est connue sous le nom de reconnaissance des entités nommées (REN).

En se basant sur la manière d'« apprendre » aux ordinateurs à analyser les

langues humaines, on peut classer les méthodes de TALN en deux catégories : les approches stochastiques et les approches linguistiques.

Les approches stochastiques modélisent une langue comme une distribution de probabilité sur des séquences de mots [87]. Ensuite, la machine peut apprendre à classer les points de données en fonction de leur similarité si la modélisation extrait des caractéristiques représentatives du texte.

Les approches linguistiques utilisent des règles de grammaire et des dictionnaires pour décrire les langues humaines pour les ordinateurs. L'objectif est alors de faire correspondre le texte aux règles. Pour réaliser nos tâches de classification spécifiques à un domaine, nous généralisons les règles de grammaire et les dictionnaires pour obtenir des connaissances complètes sur la santé des plantes, comme une liste des maladies connues d'une variété de pommes ou une liste des conditions environnementales qui favorisent la germination des spores fongiques.

Ce chapitre dresse un panorama des ressources existantes pour l'intégration des données textuelles en agriculture afin d'évaluer la faisabilité des approches stochastiques et des approches linguistiques. Concernant les graphe de connaissances existant dans le domaine de la santé des plantes, les travaux examinés présentent des approches pour la description des données en utilisant des ontologies. Cependant, ils ne traitent pas du mappage automatique de sources de données hétérogènes pour construire de tels graphes de connaissances dans le domaine de la santé des plantes. Les technologies TALN, notamment les modèles de langage pré-entraînés (PLM), semblent être prometteuses dans d'autres domaines. Mais le fonctionnement du PLM reste encore à expliquer. Notre besoin le plus essentiel pour l'extraction d'informations dans un contexte de surveillance de la santé des plantes est de fournir des connaissances formelles réutilisables à l'ordinateur et de faciliter les tâches de classification. Nous proposons de construire le PLM comme une base de connaissances implicite, qui donne des intuitions à la machine pour extraire des informations et alimenter des connaissances explicites comme des graphes de connaissances.

Chapitre 3. Informativité dans Twitter

L'exploration de données dans les médias sociaux a été largement appliquée dans différents domaines pour surveiller et mesurer les phénomènes sociaux, tels que l'analyse de l'opinion à l'égard d'événements populaires, l'analyse des sentiments d'une population, la détection des effets secondaires précoces des médicaments et la détection des tremblements de terre. Les médias sociaux incitent les gens à partager des informations dans des environnements ouverts. Face aux nouveaux verrous techniques et à la perte des connaissances locales en agriculture, il est urgent de rétablir une agriculture de précision centrée sur l'agriculteur. La question

est de savoir si les médias sociaux comme Twitter peuvent aider les agriculteurs à partager leurs observations en vue de la constitution de connaissances agricoles et d'outils de surveillance.

Ce chapitre présente les différentes étapes du développement de notre preuve de concept lors de l'initialisation de la collaboration avec les experts agronomes. Au stade du démarrage à froid, nous n'avons pas de sujets de recherche concrets, comme un ravageur spécifique, et les experts agronomes ne peuvent pas imaginer les informations extractibles de Twitter ni la granularité de ces informations. Nous commençons donc à démontrer la puissance du traitement automatique des langues avec des approches non supervisées, simples, explicables et rapides à mettre en place, notamment (1) le comptage des mots-clés populaires par mois et (2) le regroupement des sujets relatifs à la santé des plantes sur Twitter avec le modèle BoW. Puis les experts déterminent en retour différents cas d'utilisation qualitatifs et quantitatifs. Après, nous collectons, nettoyons et pré-analysons les tweets pour chaque cas d'utilisation et demandons aux experts du domaine si ces tweets répondent à leurs besoins. Ensuite, les experts examinent et étiquettent les tweets utiles pour eux : les observations des individus sur les risques naturels. Enfin, nous construisons des classificateurs avec des modèles de langage pré-entraînés pour valider que ces tweets sont « identifiables » avec les technologies existantes. Notre expérience montre que le crowdsensing sur Twitter ne remplace pas les autres paradigmes d'épidémiosurveillance, mais constitue une source d'information complémentaire. L'objectif du crowdsensing sur Twitter est de détecter les signaux faibles plutôt que de quantifier la gravité d'un problème par la fréquence des mentions. Il peut être intéressant de croiser ces informations avec d'autres sources de données. Notre approche est facilement adaptable à d'autres recherches multidisciplinaires axées sur le TALN, comme les sciences sociales computationnelles.

Chapitre 4. ChouBERT

À l'ère de la numérisation, les différents acteurs de l'agriculture produisent de nombreuses données. Ces données contiennent des connaissances historiques déjà latentes dans le domaine. Ces connaissances permettent d'étudier précisément les risques naturels sous des aspects globaux ou locaux, puis d'améliorer les tâches de prévention des risques et d'augmenter les rendements. En particulier, les Bulletin de Santé du Végétal (BSV) fournissent des informations sur les étapes de développement des risques phytosanitaires dans la production agricole. Cependant, comme nous l'avons examiné dans 2, les connaissances dans les BSV sont encore loin d'être interprétables par un ordinateur. Les tweets contiennent aussi des informations sur les aléas naturels en agriculture, moins formelles que le BSV, mais pertinentes et généralement en temps réel.

Compte tenu de la nature de ces publications, il n'est pas simple d'exploiter efficacement les informations qu'elles contiennent, et encore moins de le faire automatiquement et de relier ces données à celles provenant d'autres sources telles que des capteurs ou d'autres systèmes d'information. Pour manipuler, traiter et rendre ces données consultables, il est nécessaire de commencer par classifier automatiquement leur contenu textuel.

Ce chapitre propose d'apprendre automatiquement des connaissances sur les problèmes de santé des plantes à partir des BSV et d'appliquer ces connaissances pour améliorer l'extraction d'informations et la classification des documents à partir de sources de données textuelles hétérogènes. Les récents travaux sur le « Bidirectional Encoder Representations from Transformers » (BERT) [42] ont montré des améliorations importantes dans le domaine du TALN ; par exemple, l'efficacité des modèles BERT finement réglés pour la classification multi-étiquettes de tweets a été prouvée dans la surveillance des catastrophes [197].

Nous avons construit ChouBERT par continuer le pré-entraînement de CamemBERT sur les BSV et les tweets afin d'augmenter la représentation contextualisée des tweets pour détecter les observations. Nous mettons en évidence la généralisabilité de la représentation de ChouBERT sur des dangers non vus pendant l'entraînement de classifieur. Ensuite, nos expériences de détection d'entités de risques naturels prouvent que le pré-entraînement de ChouBERT peut également profiter aux tâches de TALN au niveau des jetons dans le domaine phytosanitaire. Nous généralisons notre approche pour améliorer le crowdsensing basé sur le contenu textuel des tweets en suivant les étapes suivantes : premièrement, la collecte d'un ensemble initial de tweets en utilisant des mots-clés ; deuxièmement, l'étiquetage manuel d'un petit ensemble de tweets ; troisièmement, le pré-entraînement de modèles de langage en utilisant des documents du domaine et des tweets ; enfin, la construction d'applications NLP avec l'ensemble étiqueté et le modèle de langage adapté au domaine. C'est ainsi que nous utilisons ChouBERT pour intégrer les informations du domaine contenues dans les BSV et les tweets et convertir ces informations en « savoir-faire » implicite pour guider l'acquisition de connaissances explicites.

Chapitre 5. ChouBERT + GAN-BERT

Les PLM suggèrent un paradigme d'ingénierie de fonction objectif pour le TALN : a) pré-entraînement des modèles de langage pour extraire les caractéristiques contextualisées du texte, suivi par b) un réglage fin avec des fonctions objectifs spécifiques à la tâche [99]. Nous avons présenté ChouBERT dans la section précédente, qui s'est avéré être une technologie prometteuse pour la détection des risques pour la santé des végétaux sur les médias sociaux. Cependant, nous sommes

toujours confrontés au manque de données étiquetées suffisantes pour valider la capacité de ChouBERT sur d'autres objectifs de classification de texte, tels que la détection de la perte potentielle de rendement ou l'inférence en langage naturel (NLI). GAN-BERT étend le réglage fin avec des données non étiquetées dans un cadre génératif antagoniste et obtient de meilleures performances dans plusieurs tâches de classification de textes lorsque le jeu de donnée étiqueté est relativement petit. Dans ce chapitre, nous étudions la combinaison de GAN-BERT et des modèles de langage pré-entraîné sur un corpus domaine-spécifique (ChouBERT). Nous discuterons des points suivants : 1) La combinaison améliore-t-elle la tâche de classification pour la détection des aléas susceptibles de menacer les plantes ? 2) Quel est l'impact du modèle pré-entraîné de type BERT sur l'entraînement dans un contexte GAN-BERT ? Nos résultats expérimentaux montrent que la combinaison de ChouBERT et de GAN-BERT peut encore bénéficier de la généralisabilité de ChouBERT pour classer des aléas non vus. Cependant, l'entraînement de telles architectures GAN pourrait également souffrir d'instabilités supplémentaires par rapport à l'utilisation de GAN-BERT avec des modèles de langage général comme CamemBERT. Ces instabilités ouvrent une autre voie d'évaluation des modèles de langage pré-entraînés et appellent à des études futures avec d'autres corpus et PLMs.

Chapitre 6. Conclusions et perspectives

Conclusions

Cette thèse propose : (1) d'utiliser Twitter comme une plateforme ouverte de crowdsensing pour acquérir les perceptions des individus sur la santé des cultures afin que nous puissions inclure la participation des agriculteurs dans la reconstruction des connaissances agricoles (2) de se servir des modèles de langage pré-entraînés comme une base de connaissances implicite et spécifique au domaine qui intègre des textes hétérogènes et supporte l'extraction d'informations à partir de textes. Stimulée par les progrès du TALN et des technologies d'apprentissage automatique, notre approche facilite l'extraction d'informations à partir de textes lorsque les ressources sémantiques sont encore limitées, et aide à alimenter les graphes de connaissances explicites. Plus important encore, notre application sur les données Twitter implique davantage de contributions humaines à l'acquisition de connaissances, ouvrant ainsi la voie à l'intégration de l'intelligence des agriculteurs dans les paradigmes de détection intelligente de la santé des plantes.

Perspectives

Nous organisons le travail futur en deux aspects : l'industrialisation et les autres directions de recherche.

Pour une utilisation industrielle, nos modèles ChouBERT suivent l'implémentation PyTorch des transformateurs et sont prêts à être déployés. Par exemple, nous prototypons un tableau de bord pour surveiller la santé des plantes sur Twitter. Sur un serveur local, nous extrayons les tweets, les annotons avec des vocabulaires contrôlés, les classons avec le classificateur de texte de ChouBERT et poussons les tweets classés vers un classeur Google Sheet. Nous visualisons le classeur dans une application Google Data Studio. Google Sheets nous permet d'accorder des rôles de lecture/écriture à différents utilisateurs. Dans un contexte d'apprentissage actif, les utilisateurs proposent de corriger la classification dans le classeur et déclenchent un nouvel entraînement du classificateur de tweet basé sur ChouBERT. En ce qui concerne le classificateur REN de ChouBERT, nous suggérons de l'utiliser directement ou de l'intégrer dans un plugin pour tout annotateur open-source qui supporte l'apprentissage actif comme INCEpTION [91].

En ce qui concerne les directions de recherche futures, tout d'abord, ChouBERT ouvre la voie à de nombreux sujets de recherche passionnants en « BERTologie » [142], tels que l'exploration des connaissances à l'intérieur du modèle de langage [130, 140, 187], le few-shot learning [56, 166], la construction de modèles multilingues pour améliorer les performances du classifieur avec des ressources limitées et des données étiquetées non équilibrées [117], ou l'application de ChouBERT à d'autres tâches d'extraction d'informations comme l'établissement de liens entre entités et l'extraction de relations. Nous pouvons également étudier l'optimisation du modèle, en essayant par exemple des architectures plus petites ou en appliquant la technique de distillation des connaissances [154]. Deuxièmement, comme nous modélisons ChouBERT comme une base de connaissances implicite formelle, les graphes de connaissances dernièrement développés (bases de connaissances explicites formelles), devraient « s'internaliser » dans les modèles de langage. Un exemple direct est de guider l'extraction d'informations avec les règles de l'ontologie (extraction d'informations basée sur l'ontologie) [43, 192] et d'alimenter les textes les plus pertinents pour le pré-entraînement et le réglage fin de ChouBERT. Troisièmement, l'application de l'exploration de texte à d'autres types de textes. Bien que nos expériences soient limitées au contenu textuel des tweets et des BSV, notre solution peut prendre n'importe quel texte en français comme données d'entrée. Une direction intéressante est la reconstruction et la recontextualisation des connaissances traditionnelles en agriculture via l'extraction de proverbes et de dictons d'agriculteurs.

Chapter 1

Introduction

1.1 Motivation

Recent advances in Information and Communication Technology (ICT) aim at tackling some of the most important challenges in agriculture we face today [35]. Researchers have applied a wide range of technologies to tackle some specific goals. Among these goals: climate prediction in agriculture using simulation models [63], making the production of certain types of grains more efficient and effective with computer vision and Artificial Intelligence [124], soil assessment with drones [171], and the IoT paradigm when connected devices such as sensors capture real-time data at the field level and that, combined with Cloud Computing, can be used to monitor agricultural components such as soil, plants, animals, weather and other environmental conditions [123]. Using such ICTs to improve farming processes is known as *smart farming* [194].

Facing the challenge of the growing population and changing dietary habits, precision agriculture emerges to increase food production sustainability. Indeed, food production sustainability is part of the "zero hunger" goal of the 2030 Agenda for Sustainable Development of the United Nations [96]. Phytosanitary issues, including (a) biotic stresses such as weeds, insect pests, animals, or pathogenic agents harmful to plants or plant products, and (b) abiotic stresses such as floods, drought, extremes in temperature, can cause a loss in food production. An essential subject in precision agriculture is improving risk prevention tasks and measuring

natural hazards within global and local aspects through real-time monitoring.

On the other side, farmers have been practicing the "modern farming" technologies since the third agricultural revolution in the late 1960s, which involve chemical fertilizers, high-yielding varieties, mechanization and irrigation. Reducing the usage and the impact of pesticides to support the sustainable development of production agriculture could re-expose farmers to the incertitude of the yield, the non-controllable factors of the production and complex phenomenons of which they do not have stabilized knowledge [133]. Traditional holistic approaches may help. Such approaches, still driven by local knowledge and hybridization of empirical evidence, are being ignored for the "inefficacy" in the short term. Using new technologies may introduce new socio-technology lock-ins [74], reinforcing dominant decision procedures. Technological evolution and climate changes demand knowledge transition in agriculture.

1.1.1 Towards textual data integration

A survey about the obstacles to applying Big Data in agriculture [191] mentions the errors in the data, the inaccessibility due to the volume of the data and the lack of communications bandwidth in rural areas, the incompatibility among different data stores and process tools, and the unusability due to the heterogeneity of data. Indeed, in the context of smart farming, **IoT devices** themselves are both data producers and data consumers and they produce *highly-structured data*. Important information related to agriculture can also come from different sources such as official periodic reports and journals like the French Plants Health Bulletins (BSV, for its name in French *Bulletin de Santé du Végétal*),[1] social media such as Twitter and farmers experiences.

The goal of the BSV is to: i) present a report of crop health, including their stages of development, observations of pests and diseases, and the presence of symptoms related to them; and ii) provide an evaluation of the phytosanitary risk, according to the periods of crop sensitivity and the pest and disease thresholds. The BSV and other formal reports are *semi-structured data*. The knowledge production of BSV is as follows:

[1] https://agriculture.gouv.fr/bulletins-de-sante-du-vegetal

1. The observation network Epiphyt[2] collect and centralise following agronomic observations made throughout France by regional monitoring networks involving 400 observers in 36 partner networks to build a national database.

2. The experts build statistical models [111] to analyze the collected data, extract critical information and make predictions.

3. Report generation processes compile the information and predictions into alerts in natural language or graphics to constitute the contents of the BSV. In this step, machine-interpretable knowledge is transformed to human-understandable data.

Thus, BSVs contain valuable agricultural knowledge in France spanning decades.

Twitter -or any other social media- can be used as a platform for knowledge exchange about sustainable soil management [113], and it can also help the public to understand agricultural issues and support risk and crisis communication in agriculture [3]. We can also acquire **Farmer experiences** (aka Old farming practices or ancestral knowledge) through interviews and participatory processes. Social media posts and farmer experiences are *non-structured data*. We will further express the value of social media in the next section.

Figure 1.1 illustrates how this heterogeneous data coming from different sources may look like for farmers: information is not always explicit or timely. Natural language processing (NLP) and knowledge graphs are technologies for data integration, information extraction and knowledge reconstruction. We review NLP and knowledge graph in agriculture in Chapter 2.

1.1.2 Monitoring Plant Health on Social media

[55] classifies existing real-time monitoring technologies of natural hazards into two categories: (i) indirect monitoring by analyzing environment parameters produced by sensor networks and Internet of Things (IoT) devices to infer the probability of phytosanitary risks [119]; and (ii), direct monitoring by processing images [134]. However, current precision agriculture technologies favor large-scale mono-culture

[2]https://agroedieurope.fr/wp-content/uploads/fiche-projet-epiphyt-fr.pdf

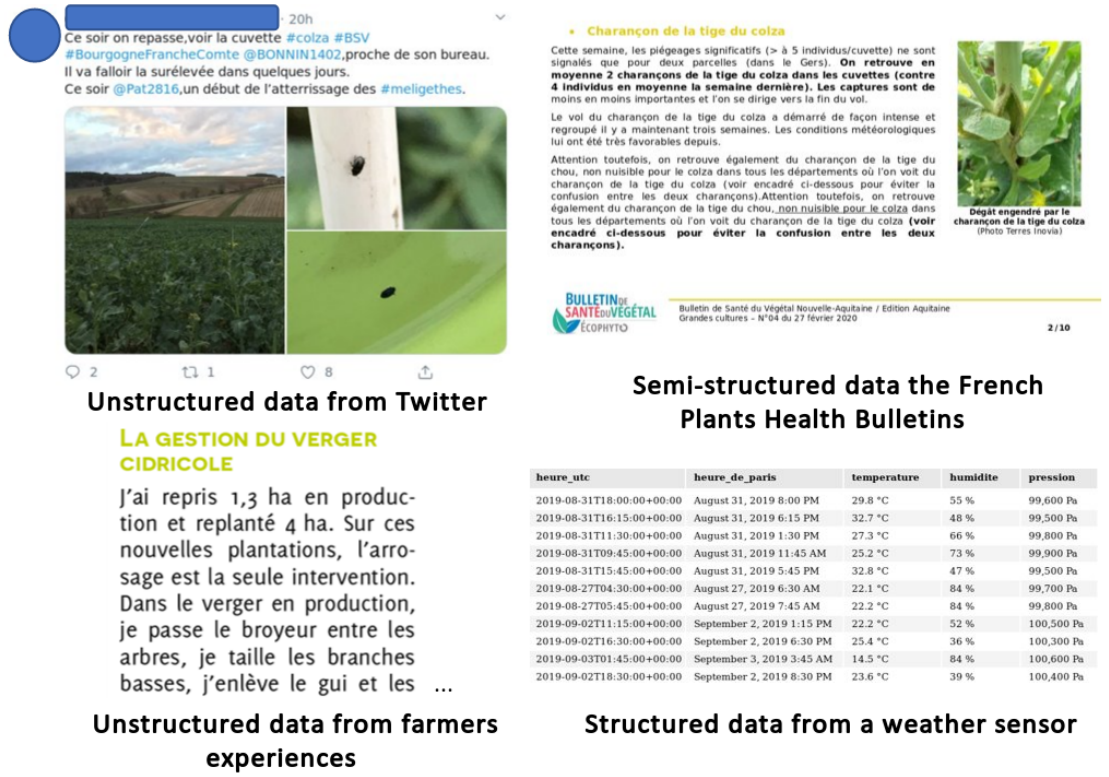


Figure 1.1: Heterogeneous sources of agricultural data: *non-structured data* from Twitter and from farmers experiences www.bio-centre.org; *semi-structured data* from The French Plants Health Bulletins; and *structured data* from a weather sensor from www.data.gouv.fr.

practices that are unsustainable and economically risky for farmers [64]. Moreover, according to the Food and Agriculture Organization of the United Nations, farms of less than 2 hectares accounted for 84 per cent of all farms worldwide in 2019, and most of these small farms are family farms [101]. Thus, we suggest that current observation data from precision agriculture cannot represent all forms of farms, especially small farms. Recently, facing the loss of legitimacy and the vanishing of local traditional knowledge, how to encourage the participation of farmers to share their knowledge and observations is drawing the attention of researchers [84, 89]. As [71] points out, local farmer knowledge relies on social processes for knowledge exchange. Still, the reducing number of farmers and the individualism weaken the local ties of blood and neighbourliness for knowledge acquisition. The diversification of professions in agricultural domain also destabilizes traditional

structures of professional sociability [167].

Social media like Twitter's role in farmer-to-farmer and farmer-to-rural-profession knowledge exchange is increasing. It suggests that the use of Twitter among rural professionals and farmers is well evolved with open participation, collaboration (retweeting) and fuller engagement (asking questions, providing answers or replies) dominating one-way messaging (new/ original tweets) [131]. Following the *social sensing* paradigm [186], individuals -whether they are farmers or not- have more and more connectivity to information while on the move, at the field level. Each individual can become a broadcaster of information. In this sense, real-time hazard information is published in social networks such as Twitter. Indeed, Twitter enables farmers to exchange experiences, subscribe to topics of interest using hashtags, and share real-time information about natural hazards. Compared to paid applications, information on Twitter, presented in the form of text, image, sound, video or a mixture of the above, is more accessible to the public but less formalized or structured. More and more farmers get involved in online Twitter communities by adding hashtags such as #AgriChatUK (http://www.agrichatuk.org) or #FrAgTw (https://franceagritwittos.com), to their posts on Twitter [41]. Thus, we can consider Twitter an open tool for farmer-to-farmer knowledge exchange.

1.2 Main contributions

The data–information–knowledge (DIK) hierarchy in [150] defines data, information, knowledge and their transformation processes as :

- Data are observed values of objects, events and their environment. Data can be processed by computers.
- Information is contextualized, and provides answers to who, what, where and when questions. Information can be inferred from data.
- Knowledge is know-how, representing the ability to understand information and then to transform information into judgments, opinions, predictions and decisions.

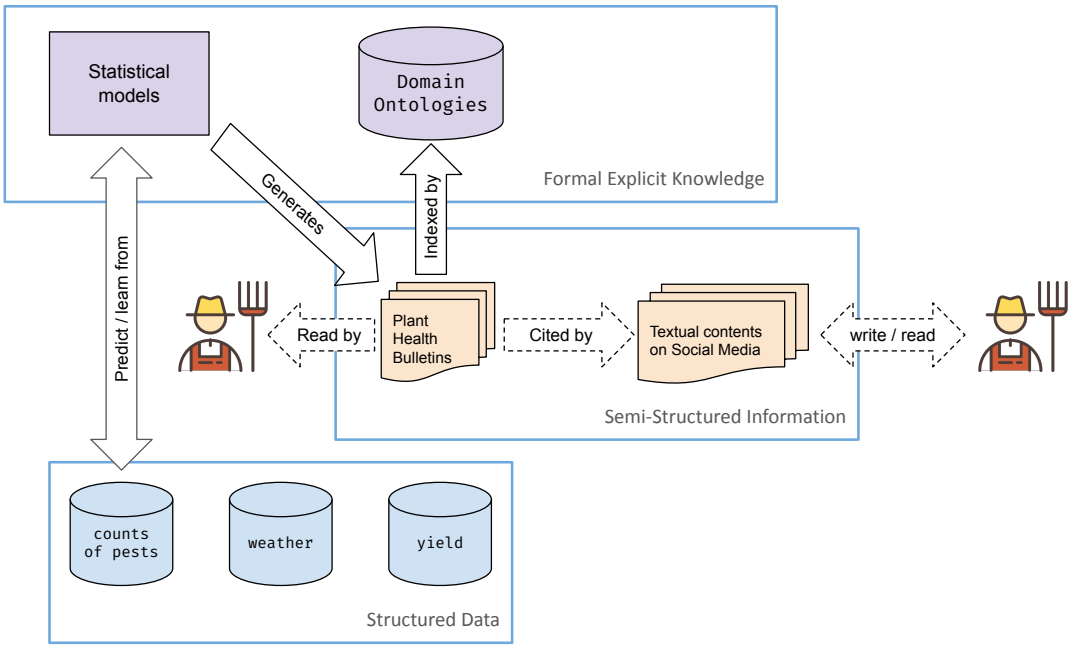


Figure 1.2: Knowledge management flow of the starting point.

Following previous studies about data-driven knowledge acquisition [143, 144], we review the current knowledge management flow with existing data sources in Figure 1.2. At this point, structured data are centralized and processed by statistical models to produce predictions. The information is then used to generate the textual content of BSV. Farmers can either contribute raw data or consume the processed information in the BSV. For computers, the processable knowledge is the statistical models and domain ontologies [148] that index the date, region, and some of the crops in the BSV. The manual development of the domain ontology limits the indexation of the BSV. In parallel, though some of the tweets mention the content of some BSV, the farmers' knowledge exchanges on social media remain untransmitted to the formal knowledge base.

Figure 1.3 illustrates the contribution of this thesis. We propose:

- To leverage Twitter as an open plant health monitoring platform to involve farmers' participation in agricultural knowledge acquisition.
- A pretrained language model ChouBERT as a formal but implicit knowledge base. This knowledge base can "automatically learn" from BSV and help to

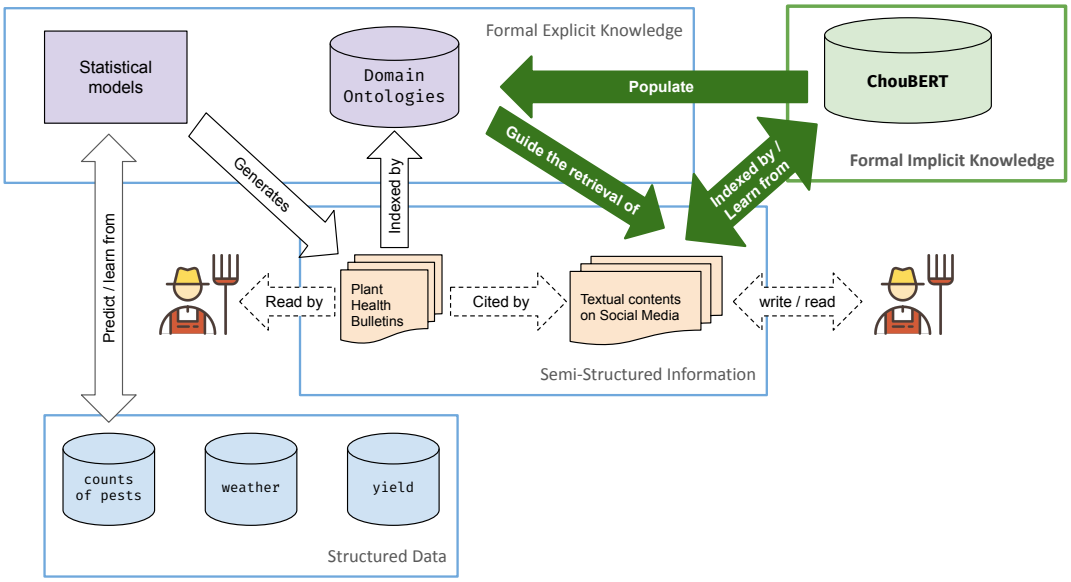


Figure 1.3: Vivace contribution to the knowledge management flow.

extract information from other textual data like tweets.

Our idea is to involve more individual knowledge and reinforce the communication between domain practitioners to reunite research organizations, technical institutes, private companies and farmers and improve the vulgarization. This thesis addresses the gap by proposing using text mining technologies to extract information from social media and to enforce plant health knowledge acquisition.

1.3 Thesis outline

We organize the thesis as follows:

- In Chapter 2, entitled "Preliminary : Background and Literature Review", we explain basic concepts related to text mining in agriculture. We cover document classification methods, text representations and existing knowledge graphs.
- In Chapter 3, entitled "Informativeness In Twitter Textual Contents For Farmer-centric Plant Health Monitoring", we develop several scenarios to col-

lect tweets, then we applied different natural language processing techniques to measure their informativeness as a source for phytosanitary monitoring.

- In Chapter 4, entitled "ChouBERT: Deep Learning for Domain-specific Information Extraction" we first validate BERT-like language model's capacity of classifying text in plant health domain. We present our pre-train language model ChouBERT as an implicit knowledge base in plant health domain. We prove ChouBERT's capacity with two supervised machine learning tasks: (a) classifying farmers' observations from other tweets and (b) annotating diseases and insect names in tweets.

- In Chapter 5, entitled "Combining GAN-BERT setup and ChouBERT: Semi-supervised Learning for Low-resource Text Classification", we explore ChouBERT with semi-supervised learning to tackle the lack of labeled data in the plant health domain.

- Chapter 4 is dedicated to the conclusions and perspectives of this work.

Chapter 2 and the first two sections of Chapter 4 refer to articles published in the scope of this thesis. Chapter 5 and the last section of Chapter 4 refer to under review articles.

1.4 Publications

1. [75]: Shufan Jiang, Rafael Angarita, Raja Chiky, Stephane Cormier, Francis Rousseaux. Towards the Integration of Agricultural Data From Heterogeneous Sources: Perspectives for the French Agricultural Context Using Semantic Technologies. Advanced Information Systems Engineering Workshops (CAiSE), 2020, Grenoble, France.

2. [80]: Shufan Jiang, Rafael Angarita, Stéphane Cormier, Julien Orensanz, Francis Rousseaux. Informativeness In Twitter Textual Contents For Farmer-centric Plant Health Monitoring. ICPRAI - 3rd International Conference on Pattern Recognition and Artificial Intelligence, Jun 2022, Paris, France.

3. [76]: Shufan Jiang, Rafael Angarita, Stéphane Cormier, Julien Orensanz, Francis Rousseaux. ChouBERT: Pre-training French Language Model for Crowdsensing with Tweets in Phytosanitary Context. International Conference on Research Challenges in Information Science (RCIS), 2022, Barcelona, Spain.

4. [79]: Shufan Jiang, Rafael Angarita, Stéphane Cormier, Julien Orensanz, Francis Rousseaux. Informativeness In Twitter Textual Contents For Farmer-centric Pest Monitoring. In: Decision Making Using AI in Energy and Sustainability. Turkey, 2022 (To appear).

5. [82]: Shufan Jiang, Rafael Angarita, Stéphane Cormier, Francis Rousseaux. Named Entity Recognition For Monitoring Plant Health Threats in Tweets: A ChouBERT Approach. In: 6th International Conference on Universal Village (IEEE UV). Boston, USA, 2022 (To appear).

6. [77]: Shufan Jiang, Rafael Angarita, Stephane Cormier, Francis Rousseaux. Fine-tuning BERT-based models for Plant Health Bulletin Classification. Technology and Environment Workshop, 2021, Montpellier, France.

7. [83]: Shufan Jiang, Rafael Angarita, Raja Chiky, Stephane Cormier, Francis Rousseaux. Vers la Reconstruction des Connaissances Agricoles : Perspectives de Détection des Risques Naturels à partir de Sources de Données Hétérogènes. Extraction et Gestion des Connaissances (EGC) Jan 2021, Montpellier, France.

8. [81]: Shufan Jiang, Rafael Angarita, Stéphane Cormier, Julien Orensanz, Francis Rousseaux. Informativité dans les Contenus Textuels Twitter pour la Phytosurveillance Centrée sur l'observation des Agriculteurs. Extraction et Gestion des Connaissances (EGC), Jan 2022, Blois, France.

Chapter 2

State of the Art

Natural language processing (NLP), or computational linguistics, is a sub-field of computer science, linguistics and artificial intelligence that explores how computers can be employed to understand, manipulate and produce human language content [33, 66]. Real-world applications of NLP include spam detection, machine translation, mining social media for monitoring environmental or social issues, information extraction from electronic health records, etc. In the context of detecting plant health issues from posts on social media, this thesis considers two NLP tasks:

- Text-level classification, such as distinguishing observations about natural hazards from other information or predicting risk based on a given textual description. The authors of [44] resume text classification as "an information retrieval (IR) process that groups documents intro predefined classes." The IR process helps filter irrelevant data with large document collections.

- Token-level classification, which aims to classify one or multiple words into predefined classes. In our case, we are interested in the pathogens, the plant traits, the developing stages of insect pests, the impacted species names and the locations. The classification of interested entities is known as named entity recognition and classification (NERC).

Based on how to "teach" computers to parse human languages, we can classify NLP methods into two categories: stochastic approaches and linguistics approaches.

Stochastic approaches model a language as a probability distribution over sequences of words [87]. Then, the machine can learn to classify the data points based on their similarity if the modelling extracts representative features in the text. We review machine learning methods and text representations in Section 2.1.

Linguistic approaches use grammar rules and dictionaries to describe human languages for computers. To achieve our domain-specific classification tasks, we generalize grammar rules and dictionaries to machine-comprehensive knowledge about plant health, like a list of known diseases of an apple variety or a list of environmental conditions that favour the germination of fungal spores. We review existing knowledge bases in Section 2.2.

2.1 Natural language processing for plant health monitoring

2.1.1 Machine learning methods for classification

Machine learning algorithms aim to recognize general rules from sample data [115]. Machine learning approaches are traditionally divided into three broad categories, depending on the nature of the "signal" or "feedback" available to the learning system:

- **Supervised learning**: The computer is presented with example inputs and their desired outputs, given by a "teacher", and the goal is to learn a general rule that maps inputs X to outputs Y. The expected output Y is called a **label**. When Y takes discrete values in a finite set, for example, determine whether the cause of the yellowing of maize leaves is a fungus or a virus, the task is called *classification*; when the outputs are continuous, for example, predict the probability of rain in Reims tomorrow morning, the task is called *regression* [32].
- **Unsupervised learning**: No labels are given to the learning algorithm, leaving it on its own to find structure in its input data, which containing

many features [59]. Unsupervised learning can be discovering interesting patterns in the input data X, such as clustering related keywords in a collection of documents.

- **Semi-supervised learning**: is an approach to machine learning that combines a small amount of labeled data with a large amount of unlabeled data during training when it is difficult or expensive to obtain labeled data. To make up for the lack of labels, semi-supervised learning try to learn from unlabeled data based on different data distribution assumptions. For example, in a classification scenario, the *cluster assumption* supposes that the data points in each class tended to form a cluster, then one could run an unsupervised clustering algorithm and propagate the label to the unlabeled points in each cluster, these later labeled data help the supervised classifier to find the boundary of each cluster more accurately [32, 180].

2.1.2 Artificial neural networks

Implementing machine learning involves creating a **model**. A model is a computer program that can process data to make predictions or find patterns. An important model for NLP is **artificial neural network** (ANN). This subsection gives general concept definitions about the neural networks related to the our experiments in Chapter 3, Chapter 4 and Chapter 5. Neural networks are networks composed of interconnected small computing units (called artificial neurons) [38, 87, 107]. Each artificial neuron takes a real-valued vector x as input, performs some computation, and yields output y. Dawson et al. [38] conclude a neuron's essential computation z as the sum of weighted input vector values (sometimes plus a scalar bias b):

$$z = \sum_i w_i x_i + b = w \cdot x + b \tag{2.1}$$

To simplify, we use θ to denote the matrix of weights and bias (w, b). Then the neural unit passes the weighted sum z to a nonlinear function f (call **activation function**) to produce the neuron's final output y. Depending on the usage, the output could be smoothly differentiable or saturated like "ON" or "OFF". In practice, the most activation functions include:

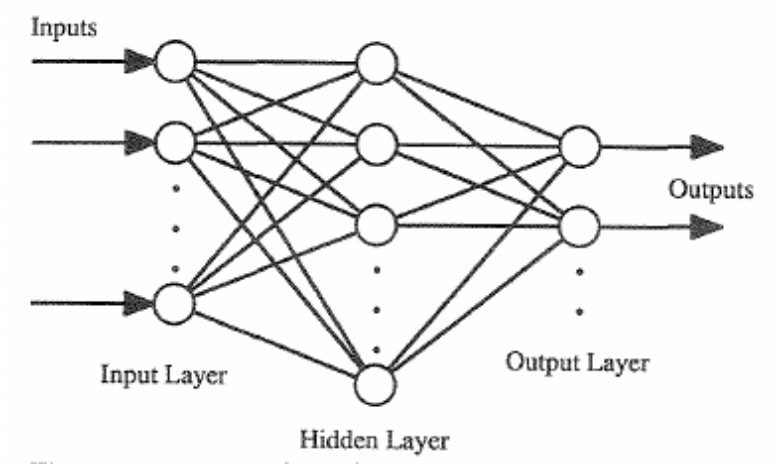


Figure 2.1: A basic overview of a feedforward neural network topology [38].

Sigmoid:

$$f(z) = \frac{1}{1 + e^{-z}} \tag{2.2}$$

Tanh:

$$f(z) = \frac{1 - e^{-z}}{1 + e^{-z}} \tag{2.3}$$

ReLu:

$$f(z) = max(0, z) \tag{2.4}$$

The neural units are arranged into layers in an ANN. An ANN is **fully-connected** when every neuron in one layer takes as input the outputs of every neuron in the next layer. When the computation in an ANN proceeds iteratively from one layer of neurons to the next, the ANN is called a **feedforward network** (see Figure 2.1). Networks with connections between neurons making cycles are known as **recurrent networks** (RNN). We will introduce other ANN architectures in the subsequent subsections.

The **training** (or **learning**) of an ANN is then adjusting the weights to a given goal. Take the feedforward network as an example. Given the expected output y for each observation x, the ANN calculates an output $\hat{y}$, and the goal is to minimize the error between y and $\hat{y}$. We use **loss functions** (also called **cost functions**) $J(\theta)$ to measure such errors. Without specification, we use the

cross-entropy loss for the experiments in this thesis:

$$J(\theta) = L_{CE}(\hat{y}, y) = -\log p(y|x) \tag{2.5}$$

where $p(y|x)$ is the probability of the correct label given the observation x. When $y \in \{0, 1\}$, then the neural network is being used as a **binary classifier**, the cross-entropy loss L_{CE} becomes binary cross-entropy loss:

$$J(\theta) = L_{CE}(\hat{y}, y) = -[y \log \hat{y} + (1 - y) \log(1 - \hat{y})] \tag{2.6}$$

When the neural network is being used as a **multi-label classifier**, then y is a vector representing the true output over K labels, the cross-entropy loss becomes:

$$J(\theta) = L_{CE}(\hat{y}, y) = -\sum_{i=1}^{K} y_i \log \hat{y}_i \tag{2.7}$$

When the vector output of the multi-label classifier allows one and only one unit to be 1 and the rest 0, the classifier is called **multi-class classifier**, such vector is call a **one-hot vector**.

Back-propagation is a strategy to adjust the weights θ by calculating the gradient (the derivative) $\nabla J(\theta)$ of the loss function with respect to each weight of the network for a (x, y) pair [59]. **Stochastic gradient descent (SGD)** is an iterative method for finding a local minimum of $J(\theta)$. SGD can be described in pseudocode as:

- Initialize the parameters θ_0 and a step size called **learning rate** η
- Loop till $J(\theta)$ reaches an approximate minimum:
  - Shuffle samples in the training set
  - For each sample (x_t, y_t) , update θ with:

$$\theta_{t+1} = \theta - \eta \nabla Loss(f(x_t; \theta), y_t) \tag{2.8}$$

Each update of θ is called a **step**, and each iteration over all the samples is called an **epoch**. At each step, we can compute the gradient with the groups

of n training sample to have a smoother convergence (called "mini-batch"). The number n of samples is called **batch size**. The parameters like learning rate and batch size, which values are used for controlling the training process, are called **hyperparameters**.

2.1.3 Text representation

Words as vectors

A simple way to project a collection of documents into vectors is to create a **document-term matrix**, which describes the term frequency (also called bag-of-words (BoW)) that occurs in a collection of documents. In a BoW model, a document is a bag of words, and these words are independent of each other.

Short for term frequency-inverse document frequency, TFIDF [85] is a formal measure of how important a term is to a document in a collection [138]. TFIDF is defined as follows. Given a collection of N documents, define f_{ij} as the frequency of a term (a word)t_i in the document d_j. Then, define the term frequency TF_{ij} of a term t_i in the document $d_j$to be f_{ij} normalized by dividing it by the frequency f_{kj} of the maximum frequency of any term in this document:

$$TF_{ij} = \frac{f_{ij}}{max_k f_{kj}} \tag{2.9}$$

The IDF of a term describes how much information the term provides. Suppose that a term t_i appears in n_i documents:

$$IDF_i = log_2(N/n_i) \tag{2.10}$$

The TFIDF score for term t_i in the document d_j is defined to be:

$$TFIDF = TF_{ij} \times IDF_i \tag{2.11}$$

The terms with the highest TFIDF scores are often the most relevant terms to represent the document's topic. TF matrix and TFIDF matrix are widely used for describing the features of the document [18].

Similar to the document-term matrix, we can also create a **term-terms matrix** to describe that term t_i and term t_j co-occur in a context. The context could be the same document, but most commonly a small window of several words around term t_i. An alternative weighting function to TFIDF in a term-term matrix can be pointwise mutual information (PMI) [50]. PMI measures how often we observe the co-occurrence of two words.

$$PMI(t_i, t_j) = \log_2 \frac{P(t_i, t_j)}{P(t_i)P(t_j)} \tag{2.12}$$

We might need to preprocess the text to reduce noise and retain only useful information in the text before applying the weighted term vectors. We categorize **text preprocessing** techniques as:

- **Noise removal**. Common noises in French Tweets can be hashtags, shortened URLs, emojis, punctuation symbols and other special characters. Depending on the purpose of the NLP task, people may filter out a list of insignificant functional words called "stop words" (e.g. "d", "le", "avoir").
- **Text normalization**, which aims to group related tokens to a root term. Such as converting the text to lower or upper case, automatically correcting potential typos, stemming and lemmatizing. In the context of detecting plant health issues, domain-specific vocabularies such as [31, 174] can help to standardize Abbreviations (e.g. "CBT du colza" to "*Charançon du bourgeon terminal* du colza") or to map varieties to their species (e.g. "golden", "gala" and "Pink Lady" are all apple varieties).

For example, consider the following tweets:

1. *"Moi dans mon MaÏs, j'ai de la pyrale, peut on parler de biodiversité ??"*
2. ''#Dégâts : Ma #parcelle de #maïs ravagée par des #chouca"
3. *"À Laval, la librairie Corneille fermée à cause d'un dégât des eaux https://t.co/qcyGB0mZdP https://t.co/Y1TyiaL2Qa"*

After the preprocessing, we obtain a vocabulary of 11 distinct words: ["biodiversité", "choucas", "corneille", "dégât", "eaux", "fermée", "librairie", "maïs", "parcelle",

“pyrale”, “ravagée”], which correspond to the 11 features in the document-term matrix in Table 2.1 and in the TFIDF matrix in Table 2.2. Each column of these tables is an 11-dimension representation of the tweet. We can see that these representations are sparse. Indeed, while increasing the size of the vocabulary and the number of documents in the corpus, the term vectors grow longer (the number of features equals the size of the vocabulary) and sparser (most elements are zero), resulting in more weights to tune in machine learning.

Table 2.1: Document-Term matrix of the 3 example tweets.

tweet	1	2	3
biodiversité	1	0	0
choucas	0	1	0
corneille	0	0	1
dégât	0	1	1
eaux	0	0	1
fermée	0	0	1
librairie	0	0	1
maïs	1	1	0
parcelle	0	1	0
pyrale	1	0	0
ravagée	0	1	0

Table 2.2: TFIDF matrix of the 3 example tweets.

tweet	1	2	3
biodiversité	0.6228	0.0000	0.0000
choucas	0.0000	0.4905	0.0000
corneille	0.0000	0.0000	0.4674
dégât	0.0000	0.3730	0.3554
eaux	0.0000	0.0000	0.4674
fermée	0.0000	0.0000	0.4674
librairie	0.0000	0.0000	0.4674
maïs	0.4736	0.3730	0.0000
parcelle	0.0000	0.4905	0.0000
pyrale	0.6228	0.0000	0.0000
ravagée	0.0000	0.4905	0.0000

Language model

Another flaw that comes with the BoW model is that there are no notion of the order of the words in a text document. Probabilistic language modeling aims to compute the probability of a sequence of words. A **language model** is a probability distribution over sequences of words [22, 87].

Without a specific downstream NLP task to evaluate the performance of a language model, we suppose that a language model with good potential gives the highest probability of a sentence in an unseen test set. Perplexity is a metric of how well a probability distribution predicts a sample and is widely used for language model evaluation. The perplexity of a language model on a test set $W = w_1, w_2, ..., w_m$ is defined as the inverse probability of the test set, normalized by the number of words [87]:

$$PP(W) = P(w_1, w_2, ..., w_m)^{\frac{1}{m}} = \sqrt[m]{\frac{1}{P(w_1, w_2, ..., w_m)}} \tag{2.13}$$

An n-gram is a contiguous sequence of n words in a given text. In an **n-gram model**, the probability of observing the i^{th} word is assumed to be approximately the conditional probability of observing it in the context of the previous $(n-1)$ words, the probability $P(w_1, ..., w_m)$ of observing the sentence of m words is approximated as

$$P(w_1, ..., w_m) = \prod_{i=1}^{m} P(w_i | w_1, ..., w_{i-1}) \approx \prod_{i=1}^{m} P(w_i | w_{i-(n-1)}, ..., w_{i-1}) \tag{2.14}$$

The BoW model in the previous subsection is unigram. The N-gram model enables us to keep some expression entire in information extraction tasks. For example, with a trigram model, we can have "pyrale du maïs" as a feature column in the Document-Term matrix. Google Book's Ngram viewer [188] is a great application of N-gram model. Given a phrase, it displays a histogram of the occurrences of the phrase in its collections of books between 1800 and 2019. For example, in Figure 2.2, all the bi-grams that terminates by "raiponce" and "Raiponce" in Google's French book corpus. In a short context of one word, we can infer that since 2000 most occurrences of the word "Raiponce" refer to Brothers Grimm's fairy

tale “Rapunzel” or its film adaptations, and that the leaf vegetable “raiponce” vanishes from French tables in the 19th century. As the number of n-grams increases

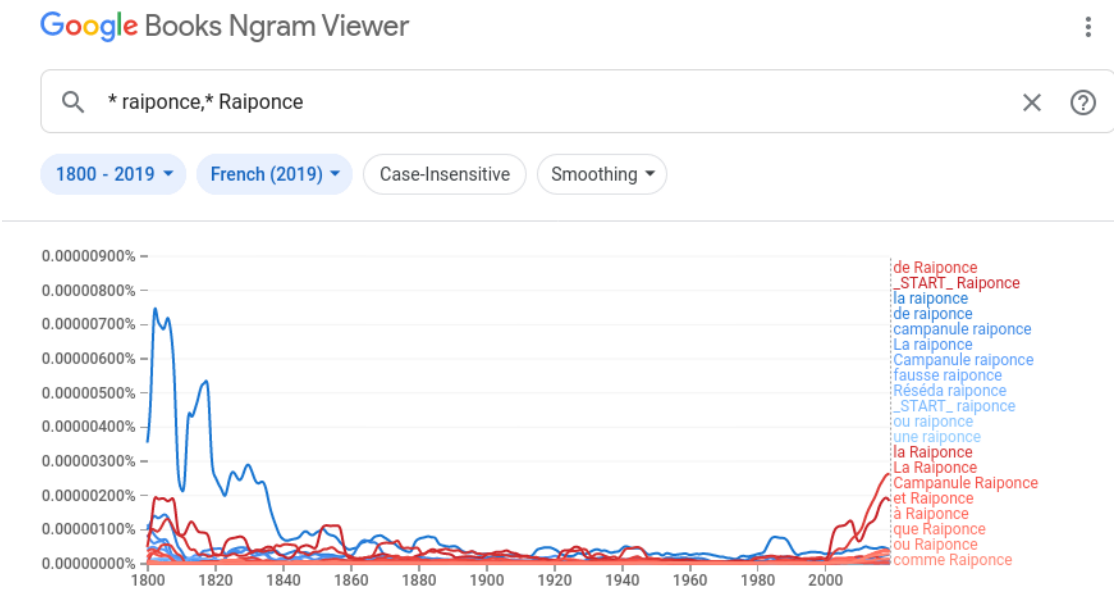


Figure 2.2: Occurences of “* raiponce” and “* Raiponce”. Google Ngram viewer allows a wildcard “*” in each query, and yields the top ten substitutions.

exponentially with the vocabulary size, causing a data sparsity problem, N-gram models can only capture a short context in practice. TFIDF weighting and N-gram models can be used to build baseline models in text mining tasks.

Static word embeddings

To tackle the curse of dimensionality and to create a notion of similarity between words, Bengio et al. [22] conceives a **neural language model** to:

1. map each word in a vocabulary to a distributed feature vector $e \in R^d$, where d is much smaller than the size of the vocabulary
2. express the joint probability function of word sequences in terms of the feature vectors of these words in the sequence, and
3. learn the word feature vectors and the parameters of that function simultaneously.

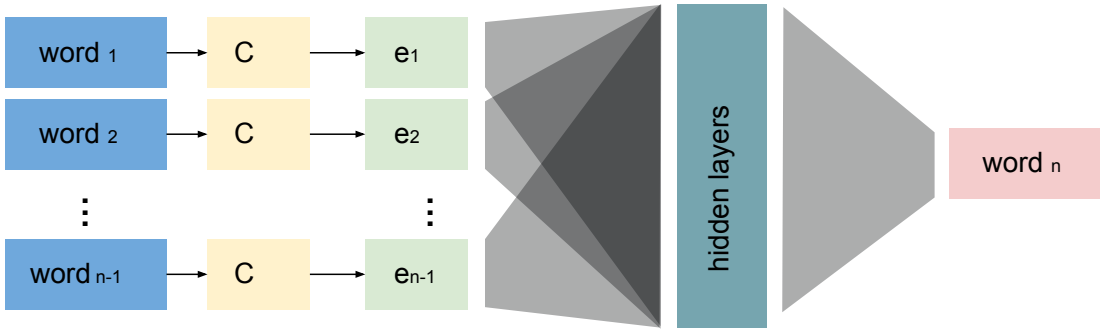


Figure 2.3: Illustration of a feedforward neural network-based language model in [22], where the matrix C maps each word in the context of $(n-1)$ words to a low dimensional feature vector e, and the hidden layers $h$estimate the probability of each word i in the vocabulary being the nth word $P(w_n = i|w_1, ..., w_{n-1})$.

The whole architecture in Figure 2.3 is a neural network. The inputs can be the feature values of a sample of external data, such as images or documents, or they can be the outputs of other neurons. The outputs of the final output neurons of the neural net accomplish the task, such as recognizing an object in an image.

To find the output of the neuron we take the weighted sum of all the inputs, weighted by the weights of the connections from the inputs to the neuron. We add a bias term to this sum.

The low-dimensional feature vectors of the words are called **word embeddings**. The original notion of "embedding" is an embedding in mathematics, meaning one instance of some mathematical structure contained within another instance. In NLP, word embeddings refers to a real-values vectors that encodes the semantics of the words, then the embeddings of the words have similar meanings are expect to be close in the vector space [87]. The training of a whole neural languge model can be expensive, several algorithms are proposed to train the mapping matrix called C (see Figure 2.3) on large corpus and produce reusable word embeddings for diverse NLP tasks, such as Word2Vec [112], GloVe [127] and Fasttext [86]. To properly handle rare or out of vocabulary (OOV) words without involving a too large dictionary, some embedding models, such as Fasttext [27, 86] and WordPiece [195], learn vector representation of character-level n-grams instead of entire words to benefit from subword information. Besides, existing knowledge bases can also contribute to build meaningful mapping, for example, EVE [135]

proposes to generate explainable word vectors using structured information from Wikipedia.

The similarities among the pre-tained word embeddings on large corpus can suggest synonyms, alternative spellings and typographical errors of word in its vocabulary. When there are few labeled data, these substitutions can be used to paraphrase sentences to multiply the dataset, aka. *data augmentaion* [199]. However, the pairwise similarities of word representations on large and general corpus do not necessarily make sense in an agricultural context. For example, in the vector space of French Word2vec, the word "blé" (wheat) has the highest cosine similarity to "maïs" (maize). Indeed, one of the main limitations of word embeddings (word vector space models in general) is that words with multiple meanings are conflated into a single representation. In other words, polysemy and homonymy are not handled properly. A possible solution is to build domain-specific word embeddings, for example, BioWordVec [200] is a set of biomedical word embeddings that improves performance in multiple NLP tasks in the biomedical domain. However, as we aim to integrate heterogeneous textual data, we still need to process out of domain text like noises in tweets, thus we need to handle the ambiguous words. In Section 4.3, we list such words in Table 4.11.

Contextualized embeddings

Following the idea of pre-training reusable embeddings, contextualized embeddings project a word's context in the hidden layers in Figure 2.3 to disambiguate polysemous words. **Pretrained language models (PLMs)** are deep neural networks of pre-trained weights to vectorize sequences of words. Such vectorial representations obtain state-of-the-art results on NLP tasks like text classification, text clustering, question-answering and information extraction. PLMs suggest an objective engineering paradigm for NLP: language model pre-training for extracting contextualized features from text and fine-tuning with task-specific objective functions[99]. Recurrent neural networks, such as long short-term memory (LSTM) [67] and gated recurrent units(GRUs) [34], can capture contextual information among units in a neural network and revolutionized several sequence modeling and transduction protblems [19, 110]. Universal Language Model Fine-tuning

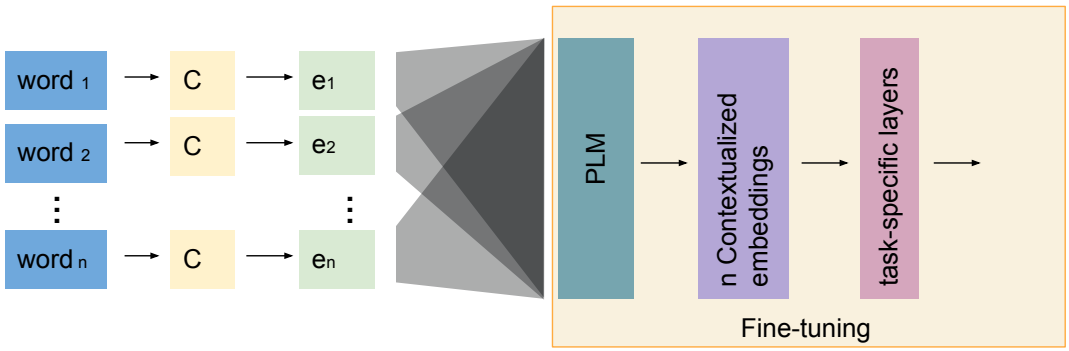


Figure 2.4: PLMs suggest an objective engineering paradigm for NLP.

(ULMFiT) [69] and Embeddings from Language Models (ELMo) [129] are deep-contextualized PLMs based on LSTM architectures. However, recurrence disfavors parallelization in the computation, and restrains representing fully bidirectional relations between phrases in different positions of the sequence. **Transformer** is "a model architecture eschewing recurrence and instead relying entirely on an **attention mechanism** to draw global dependencies between input and output" [182] . A comparison [47] between ELMo and transformer-based PLMs (BERT [73] and GPT-2 [137]) shows that contextualized word embeddings are more context-specific in higher layers on average. And Transformer architecture are more adaptable for piling up layers. Next, we will introduce more details about Transformer.

Self attention mechanism In general, attention is a technique to enhance some phrases of the input sentence by computing a soft weight to the static embedding of each word. Depending on the usage and the architecture of the ANN, there are many variants of attention [19, 182]. Transformer is built with scaled dot-product attention units called **self attention** that relate different positions of the input sequence to compute a representation of the sequence. Imitating the retrieval of a value v_i for a query q based on a key k_i in databases, self attention treats each word as a query and finds some keys that correspond to the other word of the sentence:

$$attention(q, k, v) = \sum_i similarity(q, k_i) \times v_i \tag{2.15}$$

Thus, the transformer model learns three weight matrices for each self attention unit : the query weights $W_Q \in \mathbb{R}^{d_{model} \times d_k}$, the key weights $W_K \in \mathbb{R}^{d_{model} \times d_k}$, and the value weights $W_V \in \mathbb{R}^{d_{model} \times d_v}$. $d_{model} = 512$ refers to the identical output sequence size of all the layers in the model, d_k and d_v refers to the dimensions of key vectors and query vectors. Transformer is built with blocks of h parallel attention units, thus the authors set $d_k = d_v = d_{model}/h = 64$. For the input word in position i, we multiply the word embedding x_i with these three matrices to get the query vector : $q_i = x_i W_Q$, the key vector: $k_i = x_i W_K$, and the value vector $v_i = x_i W_V$. These vectors are then packed together into matrices Q, K and V. Then the similarity s_{ij} from word i to word j is calculated by the dot product of these vectors: $s_{ij} = q_i \cdot k_j$, thus the pairwise dot products of queries and keys can be written as $Q^T K$. Then the similarity is divided by the square root of the dimension of the key vectors (the scaling factor) to have more stable gradients. A softmax function $softmax(s_i) = \exp(s_i)/\sum_j \exp(s_j)$ normalizes the scaled similarities so they are all positive and add up to 1. Finally the attention of Q,K,V becomes:

$$attention(Q, K, V) = softmax(\frac{Q^T K}{\sqrt{d_k}})V \tag{2.16}$$

One set of (W_Q, W_K, W_V) matrices is called an attention **head**. We use BertViz [183] to illustrate an attention head in Figure 2.5. We can observe that mBERT's tokenizer breaks words into wordpieces and that "symptômes" (symptom) pays most attention to itself and "maladie" (disease).

Transformer Aiming to resolve machine translation problems, early bidirectional RNN-based neural language models have an encoder-decoder architecture: the encoder project the input sentence in a language to a context vector, and the decoder is trained to predict the next word in another language given the context vector and all the previously predicted words [19]. Following this idea, transformer builds its encoder-decoder architecture (see Figure 2.6) with **multi-head attentions** by linearly project h times an attention head with a jointly learned

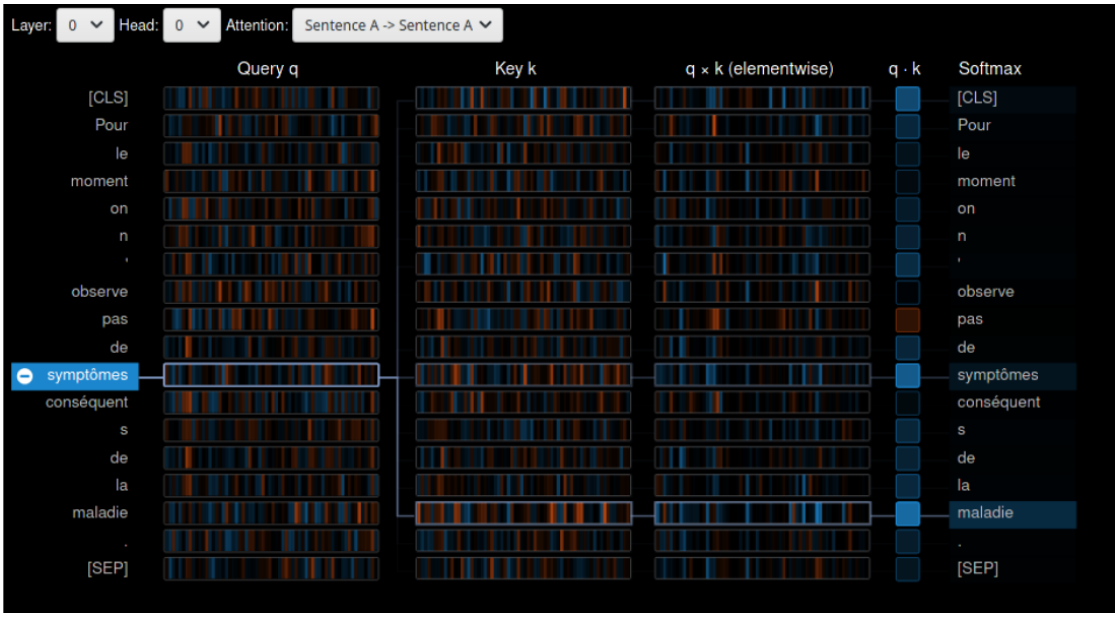


Figure 2.5: A Multilingual BERT (mBERT)'s attention head of a sentence from French Plant Bulletin.

projection matrix W^O:

$$MultiHead(Q, K, V) = Concat(head_1, head_2, \ldots, head_h)W^O \tag{2.17}$$

where $head_i = attention(QW_i^Q, KW_i^K, VW_i^V)$ and $W^o \in \mathbb{R}^{d_v \times d_{model}}$. The encoder is a stack of 6 identical layers. Each layer is composed of two sub-layers: a multi-head attention and a position-wise fully connected feed-forward network. The feed-forward networks consist of two linear transformations with a ReLU (Eq. 2.4) activation in between. The output of each sub-layer is normalized by $LayerNorm(x + Sublayer(x))$ to have 0 mean and 1 variance, which reduces "covariate shift" and makes the training converge faster. Similarly, the decoder is a stack of 6 identical layers too. Compared to an encoder layer, a decoder layer has two multi-head attention sub-layers. One is inserted over the output of the encoder stack. The other masks some values to ensure that an output value only depends on previous outputs (masked multi-head attention). To encode positional information of the token in the sequence, transformer use sine and cosine functions

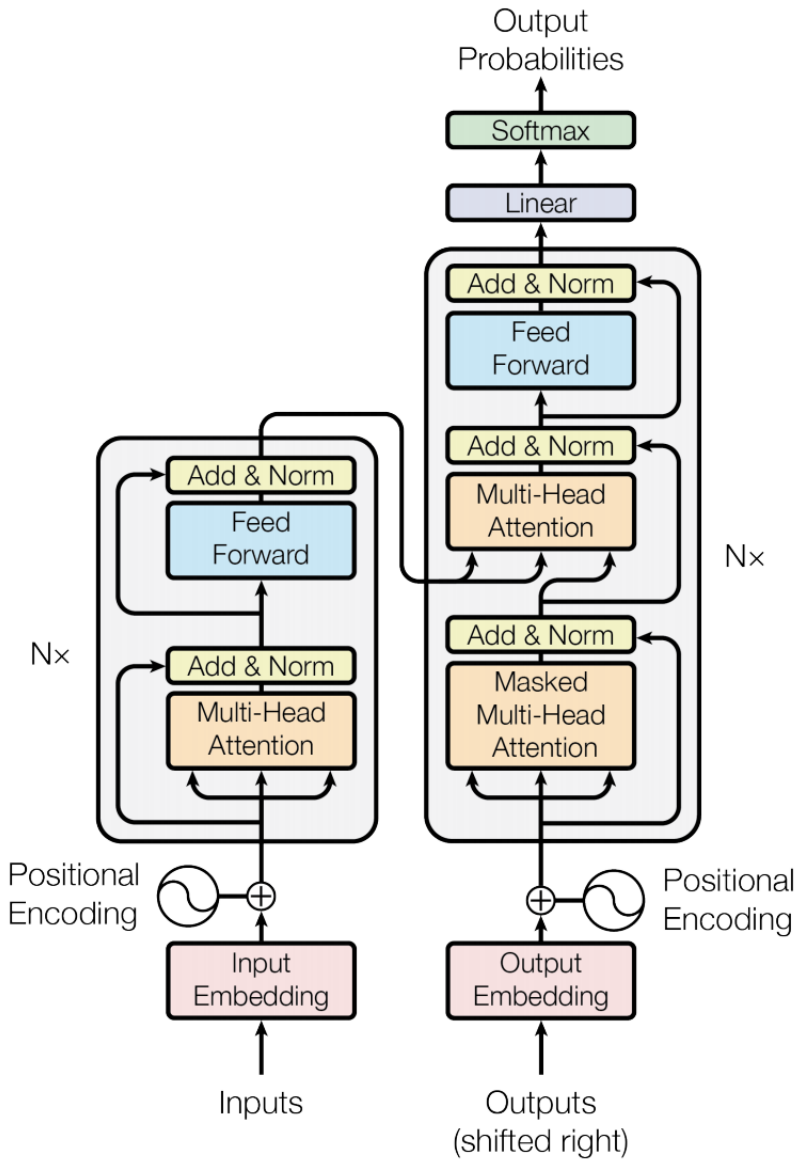


Figure 2.6: The transformer model architecture [182].

of different frequencies to encode the position i and add this positional encoding to the corresponding input embedding. Finally, the linear layer project the output produced by the decoder stacks into a large vector of the model's vocabulary size (called a logit vector), each dimension corresponding to a word. Next, the softmax

layer converts the logit vector into probabilities. Then the output is the word associated with the highest probability.

Most importantly, transformer architecture has inspired many pre-trained language models that refresh state of the art in many NLP tasks, such as:

- Bidirectional encoder representations based on transformers (BERT) [73], built with twelve stacked bidirectional encoders over the static embedding produced by WordPiece [195];
- Generative Pre-trained Transformer (GPT) [136], built with twelve left-to-right decoders;
- BART [98], built with bidirectional encoders and left-to-right decoders.

The phenomenal growth of transformer-based pretrained models also gave birth to HuggingFace[1] company, which at the same time proposes programming interfaces to facilitate the exploration of transformer-based models, a free repository for hosting and sharing models and paid services to accompany the industrialization of the models. HuggingFace builds an ecosystem to promote the development of transformer-based models.

BERT Among the transformer-based PLM mentioned above, BERT has the most significant influence in this field. BERT is pre-trained in two stages: first, a self-supervised task where masked words in a text must be retrieved by the Masked Language Model (MLM); and second, a supervised task where the model must re-find whether a sentence B is the continuation of a sentence A or not (Next-Sentence Prediction, NSP). The pre-training produces in the end 12 stacked encoders, which take a sequence of tokens as input, add a special token "[CLS]" to the beginning of the sequence, and a "[SEP]" to the end of each sentence, and calculate a fix-length vector for each token. Each vector dimension represents how much attention that token shall pay to the other tokens. The contextualized representation can then serve as an input feature to a downstream layer, for example, a classifier, and be fine-tuned with task-specific objective function and training data. In the downstream fine-tuning, the weights in the encoder lays are jointly updated.

[1] https://huggingface.co/

For the sentence-level classification task, the representation of "[CLS]" is used to represent the whole text. For the token-level classification, the representation of all the wordpieces is used. Figure 2.7 illustrates how the similarity of "[CLS]" representation changes after fine-tuning with a sentence-level classification.

Among the French varieties of BERT, CamemBERT [102] is a model based on the same architecture as BERT but trained on a French corpus with MLM only. In this manuscript, we introduce ChouBERT [76] in Chapter 4, which takes a pre-trained CamemBERT-base checkpoint and further pre-trains it with MLM over a corpus in French in the plant health domain to improve performance in detecting plant health issues on Twitter.

2.2 Knowledge bases for plant health data integration

The IoT promises the easy integration of real-time agricultural into computer-based systems. This data collected is at the field level can help to detect changes in the weather and the apparition of natural hazards such as pests and diseases. These physical objects are connected to the virtual world, allowing remote sensing and acting. People can also produce relevant agricultural data through formal reports or by publishing content in social media such as Twitter, acting as social sensors. Given the high heterogeneity of all these data, the challenge is to make sense of it in a unified way that can support the detection of natural hazards. Indeed, an important part of this work belongs to the area of *data integration and interoperability*.

2.2.1 Data interoperability

Interoperability is a characteristic of a product or system whose interfaces are completely understood to work with other products or systems, at present or in the future, in either implementation or access, without any restrictions [40, 158] defined four levels of interoperability: system, syntax, structure and semantic [198]: The *system level* concerns the exchange between different hardware components, operating systems and communication systems. For example, different IoT devices

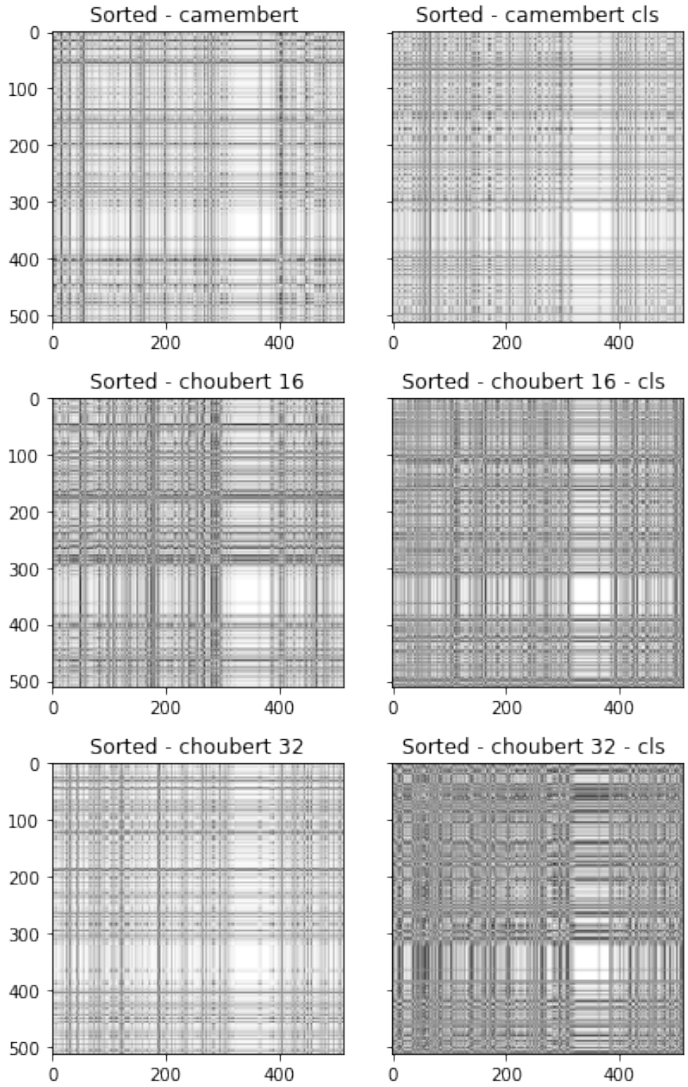


Figure 2.7: The pairwise cosine similarity between 512 data points produced by different PLMs, before and after fine-tuning with sentence-level classification. The first 256 data points are French tweets about plant health observations, the rest 256 are irrelevant tweets or noises. We use t-distributed stochastic neighbor embedding (t-SNE) [179] to project the [CLS] embedding into a 2-dimensional map.

may produce observation data at different frequencies. The *syntax level* considers data formatting including encoding, decoding and representation. For example, some applications generate data in CSV format while others read data in excel

format. The *structural level* addresses data representation heterogeneity involving data modelling constructs and schematic heterogeneity, which refers mainly to structured databases. For example, each region has its dedicated page formatting for its BSV. The *semantic level* requires that the information system understand the semantics of the user's information request and those of information sources. For example, different information systems may adopt different units of measure to describe the lengths or the weights.

This section reviews the state of the art of the integration of heterogeneous datasets from the semantic-level point of view. Goh [57] and Pánek et al.[122] identified three major causes of semantic heterogeneity: confounding conflicts, scaling and units conflicts and naming conflicts. Confounding conflicts occur when two pieces of information seem to be the same, but they are distinct. For example, "the current stage of the crop" mentioned in two different plant health reports may differ due to different tempo-spatial contexts. Scaling and unit conflicts lies in the usage of different units of measures or scales. [170] shows a case in environmental observation, where a land database uses the NTF (Paris) / Lambert zone II reference system for parcel geometries, and two other databases record the observation points on the WGS 84 one. Naming conflicts refer to differences caused by naming schemes. These conflicts are related to the usage of synonyms or homonyms. Take weeds as an example. The French plant health bulletins use the term "adventice", while in farmers' discussions on social media, it is always called "mauvaise herbe", when translated into English, "weeds" may be a common slang word for cannabis in another context.

One way to look at interoperability is by the extent two systems can interact with each other by relaying on **a common standard** [181] to which each party is willing to participate by sharing or consuming data must adhere. Following this paradigm, the authors of [156] present Breeding API (BrAPI)[2] as an API (Application Programming Interface) for exchanging plant phenotype and genotype data between crop breeding applications. Such specification allows different data providers to expose and share their data in a standard way. For example, BrAPI specifies the structure of URLs, data and error handling. The API specification is an example of interoperability at the system layer. The advantage of providing a

[2]https://brapi.docs.apiary.io/

standard API specification is that it is built on top of very well-known mechanisms and technologies, such as JSON-based HTTP requests and responses. However, each data provider handles on its own the transformation of its data to comply with the specification. Another shortcoming is that it is up to humans to interpret the semantics of the data, and there is no connection between the different data types. For example, a developer knows that "wheat" is a crop, but there is no automated way to explore other data that may be related to it, such as different crops, places, or diseases. Therefore, we need to add extra data (metadata) so that machines understand what the data is about.

2.2.2 Formalizing knowledge representation with semantic resources and technologies

Following the idea of a common format, the **Semantic Web** emerges as an extension of the World Wide Web through standards set by the World Wide Web Consortium (W3C [25]). The goal of the Semantic Web is to make data machine-readable to facilitate sharing and reusing of Internet data across applications. To achieve this goal, W3C proposes[3]:

1. the use of a Universe Resource Identifier (URI) to name resources on the web;
2. the inclusion of links to other URIs so that more data can be discovered;
3. the use graphs to describe knowledge as entities and the relations between them.

The W3C proposed Resource Description Framework (RDF)[4] as a standard format of these graphs, where knowledge is organized into triples of subject - predicate - object (see Figure 2.8).

The subject and the predicate are designated as URIs. The object can be a URI or a literal. The utilization of URI makes it possible to include all concerned resources within an RDF knowledge base. Based on RDF, several data

[3]https://www.w3.org/DesignIssues/LinkedData
[4]https://www.w3.org/RDF/

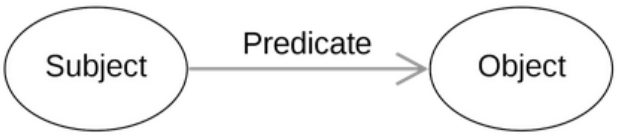


Figure 2.8: An RDF triple.

language standards are built to support more specific knowledge representation and reasoning over the semantic web. Zeng [198] lists some most important and widely applied W3C standards including: RDF Schema (RDFS)[5], Web Ontology Language (OWL)[6] and Simple Knowledge Organization Systems (SKOS)[7]. The adoption of RDF enables to query diverse RDF data sources using a standard query language and protocol : SPARQL (SPARQL Protocol and RDF Query Language)[8]. SPARQL can be describe queries with optional graph patterns along with their conjunctions and disjunctions. For example, Figure 2.9 shows a query about known pests to crops in a agronomic knowledge base. The results are the URIs of the pests and crops.

In [15, 45], different RDF-based semantic resources and technologies are used to enhance interoperability. Depending on each sub-domain or specific use case, vocabularies (controlled vocabulary, taxonomies for concept categorization, thesauri or synonym dictionaries, etc.) are used to define domain-related concepts in a standardized way. Ontologies are used to describe entities, relations, and rules of how these elements can be organized together. Vocabularies and ontologies are known as Knowledge Graphs (KG), or Knowledge Organization Systems (KOS).

The authors of [149] classify RDF-based knowledge graphs in agriculture as three types of ontologies:

- **Terminological ontologies**, including glossaries, dictionaries, controlled vocabularies, taxonomies, folksonomies, thesauri, or lexical databases. This type of ontology focus on terms and relationships. The relationships can be hierarchical links, related links or synonym links. Terminological ontolo-

[5]https://www.w3.org/TR/rdf-schema/
[6]https://www.w3.org/2001/sw/wiki/OWL
[7]https://www.w3.org/2009/08/skos-reference/skos.html
[8]https://www.w3.org/TR/rdf-sparql-query/

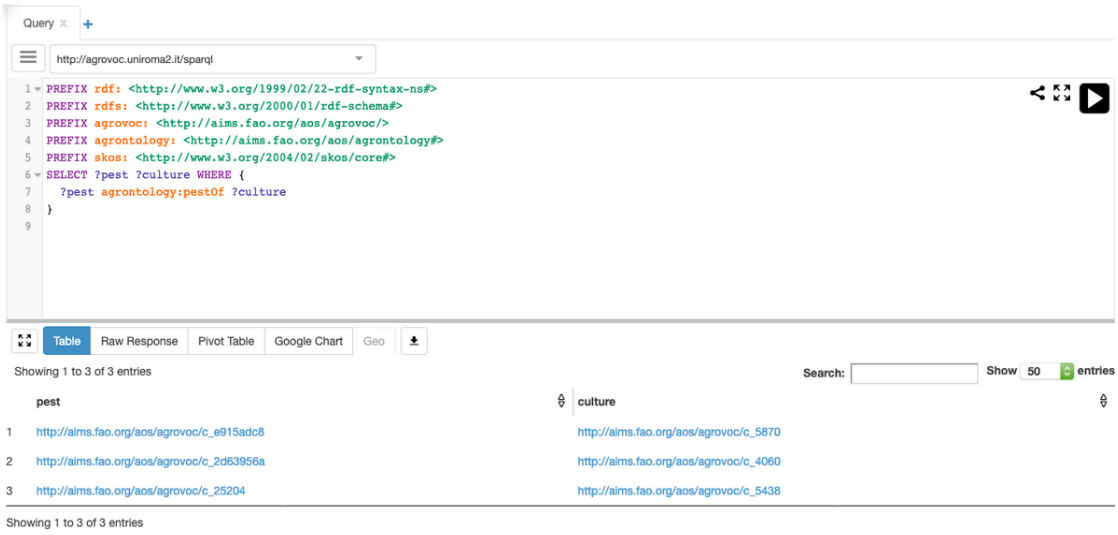


Figure 2.9: An example of using SPARQL to find pest-crop pairs in Argrovoc Thesaurus [31].

gies are used to define a domain's vocabulary, representing a terminological agreement of users' community to avoid ambiguity. RDF and SKOS are two languages specific to store terminological ontologies.

- **Data ontologies**: This type of ontology provides conceptual schemata about data storage and handling to guarantee the consistency of data exchange between different information systems. Conceptual modelling languages like Unified Modelling Language (UML) are used to define data ontologies in software and database engineering.

- **Logical ontologies**: This type of ontology uses formal logic (usually, First Order Logic or Description Logic) to define concepts, relations and rules of how concepts and relations can be combined. Formal languages used to describe logical ontology, like Description Logics (DL), Conceptual Graphs (CG), First Order Logic (FOL), and OWL, are used to describe logical ontologies.

Considering our topic about plant health and natural hazards in agriculture, we identified the following resources:

- **AGROVOC** [31] is a multilingual thesaurus about general agriculture developed by the Food and Agriculture Organization of the United Nations[9]. Till February 2021, it contains 38 400 concepts and 803 800 terms in up to 40 languages. The data set is widely used in specialized digital libraries and repositories to identify resources, index content, and translate. FAO and third-party stakeholders serve AGROVOC as a specialized tagging resource for the content organization. AGROVOC has been successfully applied for animal disease surveillance from the news [176, 177].

- **AgroRDF** [105] is an RDF Schema-based approach built for describing agricultural work. It provides data types to describe the data on work processes on the farm, including accompanying operating supplies like fertilizers, pesticides, crops and crop species, chemical substances and harvesting dates. Its purpose is to provide documentation for agricultural processes, data integration between different products and integration between different standards and vocabularies.

- **Semantic resources for integrating French Plant Health Bulletins**: The French National Institute For Agricultural Research (INRA) has been working towards publishing the bulletins as Linked Open Data [148]. Three terminological ontologies were manually constructed to publish the link of BSV and the annotations as RDF triples: **FrenchCropUsage**, a thesaurus of types of crops organized according to their destination in France. **VESPA**, an ontology for bulletin description. We remark that the regions in this ontology are out-of-date, as France reformed and merged its administrative regions in 2016. Thus, links to BSV in those regions need to be updated. **Maladies des cultures** (diseases of plants in french) [174], a vocabulary developed in SKOS for the text-mining of BSV. It covers the preferred names of 308 diseases which appear in BSV.

- **Ontologies in the plant health domain**: **CropPestO** [141] is an ontology about pest control in English and Spanish. It is aligned with AGROVOC via reusing the crops in AGROVOC, implements the isPestOf relation and

[9]http://www.fao.org

extends AGROVOC with instances and relations in a Spanish context. The **Plant Health Threats Ontology** (PHTO) [4] is a multilingual ontology for monitoring plant health threats in a media monitoring system called MedISys. This ontology models plant pests and diseases, together with other concepts related to them: affected crops, hosts, vectors and symptoms. 81 of the 117 plant threats concepts include French labels. The authors also propose to use symptom expressions to monitor unknown threats. PHTO, indeed is the nearest solution to our needs. However, the keyword-based search approach of MedISys has difficulty in reducing the ambiguity of terms. Therefore, the authors propose to assign a weight attribute as the category threshold to each threat name (scientific, common...) and use a list of "negative words" to reduce the noise in the news captured by their symptom expressions. Though the project proves the usefulness of nonofficial information sources for detecting early plant health threats, the PHTO is yet to be completed with more hand-crafted rules (e.g. adjusting the threat name weights for specific categories) and out-of-domain vocabularies (the "negative words"). Furthermore, we suggest that the weights for threat names can be obtained and optimized using stochastic approaches and machine learning.

- **EPPO Global Database** [48] provides detailed and multilingual information for more than 1700 pest species that produced and maintained by the European and Mediterranean Plant Protection Organization (EPPO). For each pest, the database gives geographical distribution, host plant and quarantine status. The EPPO also has a website for its registered members to report pests. EPPO Global database associate most of its recorded organisms with a unique and constant code (EPPO code) to avoid conflicts caused by taxonomy changes.

As decisions in the agricultural domain demand multidisciplinary knowledge, some solutions are proposed to achieve the **interoperability of knowledge graphs**:

- **Ontology alignment** or ontology matching determines correspondences between semantically related entities of overlapping knowledge graphs. A set of correspondences is also called an alignment. These correspondences

may stand for equivalence, consequence, subsumption, or disjointness [49]. The detection of such correspondences is performed via calculating similarity measures based on the comparison of terms, structural information like links between entities, entities attributes and extensions of entities like instances of classes. External resources may serve as a common context for the comparison [6]. Thiéblin et al. [168] give a detailed survey on complex ontology alignment techniques. The Ontology Alignment Evaluation Initiative (OAEL)[10] gives benchmarks for evaluating ontology matchers and benchmarking campaigns. The Global Agricultural Concept Schemec (GACS)[11] is the largest linked open data attempt in agriculture. It proposes a shared concept schema to align major large knowledge graphs such as AGROVOC of the U.N., the CAB Thesaurus by CAB International of U.K., and the U.S. National Agricultural Library (NAL) Thesaurus.

- **Ontology repositories and ontology registries** are aggregations of knowledge graphs into a larger resource [45, 198]. The difference between a repository and a registry is that a repository hosts the full content of a knowledge graph, while a registry provides the metadata for locating knowledge graph resources. These resources allow users to query the ontologies via a web interface or a SPARQL endpoint. The Crop Ontology [12], PHTO [4], CIRAD Ring [128] and Vest [13] are examples of such resources. Compared to monolithic knowledge graphs, repositories and registries enable the decentralized development of sub-domain knowledge graphs.

2.2.3 Knowledge graph construction and information extraction

The construction of a knowledge graph can be manual, automatic, or a mix of both. Manual construction requires a closed group of domain experts (curated approach) or an open group of volunteers (collaborated approach). The participation

[10] http://oaei.ontologymatching.org/
[11] http://browser.agrisemantics.org/gacs/en/
[12] https://www.cropontology.org/
[13] https://vest.agrisemantics.org/

of domain experts assures the accuracy of the knowledge, while the collaboration of an open community is easy scaling. A curated approach example is FrenchCropUsage. Domain experts developed FrenchCropUsage using competence questions [147]. A collaborated approach example is Crop Ontology (CO). CO aims to create a community of contributors interested in building standard ontologies for crop-related topic [106]. Automatic construction of a knowledge graph refers to computer-aided information extraction, entity linking and triplification. For automatic construction, there are three categories: pattern-matching-based approach, gazetteer-based approach and machine learning-based approach [160]. Depending on the structuring level of the source document, pattern-matching-based methods, such as hand-crafted rules, learned rules or regular expressions, are used for information extraction from semi-structured text such as Wikipedia infoboxes. Machine learning and natural language processing techniques are applied for information extraction from unstructured text [151]. An automatically built knowledge graph example is YAGO [164], an open-domain knowledge graphs with good scalability and accuracy.

Concerning our case, we choose automatic construction. We consider restructuring the knowledge of unstructured and semi-structured data in two steps: (1) categorize the data to make them easier to be discovered and be indexed, (2) extract information from these data depending on each category and reorganize the information in a knowledge graph. **Information Extraction (IE)** refers to the automatic extraction of implicit information from unstructured or semi-structured data sources [103]. Named Entity Recognition (NER) and Relation Extraction (RE) are two of the most important IE tasks for knowledge graph construction. A Named Entity (NE) is an entity or phrase of interest defined by a specific domain's given schema. Named Entity Recognition is the task of identifying in the text and extracting the Named Entities defined by the model. This includes detecting the existence of a NE and finding correct textual boundaries and its classification as the correct NE type. Relation Extraction is detecting and classifying relations between NEs [178]. The following NLP tools have been used for automatically extracting information to build knowledge graphs with linguistic approaches:

- Apache Unstructured Information Management Architecture (UIMA)[14] is "a component architecture and software framework implementation for the analysis of unstructured content like text, video and audio data." UIMA provides various annotators to associate a document with relevant concepts. Computer-aided Ontology Development Architecture (CODA) is an extension of UIMA for ontology learning, population, and linguistic enrichment of ontologies [52]. An essential feature of CODA is its rule-based pattern matching and transformation language, called PEARL (ProjEction of Annotations Rule Language). PEARL allows customized projection rules for RDF generation from UIMA metadata. [51, 125]

- The Stanford NLP suite [5] implements Open Information Extraction that extracts general domain relation triples from plain text. The triple represents a subject, a relation, and the object of the relation. Unfortunately, till February 2021, the Stanford OpenIE annotator is only available in English [104].

- GATE[15] offers Java libraries for tokenizing text, sentence segmentation, POS tagging, parsing, coreference resolution and terminology extraction.

- Spacy[16] is a toolbox in python that integrates pre-trained NLP pipelines components such as text cleaning, named entity recognition, part-of-speech tagging, dependency parsing, sentence segmentation, text classification, lemmatization, morphological analysis, entity linking.

- Nooj [159] is a linguistic development environment that allows the mathematical description of different linguistic phenomena at the orthographical, lexical, morphological, syntactic and semantic levels, for any natural language. NooJ's linguistic engine supports the four types of grammars of the Chomsky hierarchy to facilitate the creation of fine-grained text annotators for domain-specific information extraction.

[14] https://uima.apache.org/
[15] https://gate.ac.uk/ie/
[16] https://spacy.io/

However, comparing to the well-developed resources in English, the resources for French NLP tasks are quite limited in general multilingual NLP toolkits like Stanford NLP suite and Spacy. For example, there are only a few French stop words in the stop word list of Spacy.

2.3 Discussion and conclusions

In the context of monitoring plant health from textual data, this chapter reviews related works to evaluate the feasibility of stochastic approaches and linguistics approaches.

For stochastic approaches, we go through machine learning and text representation techniques. We find that pre-trained language models revolutionize many NLP tasks in the general and biomedical domains. However there is no French PLM in the plant health domain. When reviewing the linguistic approaches, we also remark that some tasks can be optimized with stochastic approaches.

For linguistics approaches, we focus on existing knowledge graphs. Most of the reviewed works present data descriptions using ontologies and RDF. But few existing knowledge graphs in the plant health domain describe the rules to support the reasoning for information extraction. Furthermore, the development and maintenance of these knowledge graphs still rely on the manual work of domain experts.

After all, in our context, both knowledge graphs and PLM aim to provide reusable formal knowledge to the computer and facilitate classification tasks. Therefore, our strategy is to build PLM as an implicit knowledge base, which gives intuitions to the machine to extract information and populate explicit knowledge like knowledge graphs.

Chapter 3

Informativeness In Twitter Textual Contents For Farmer-centric Plant Health Monitoring

Data mining in social media has been widely applied in different domains for monitoring and measuring social phenomena, such as opinion analysis towards popular events, sentiment analysis of a population, detecting early side effects of drugs, and earthquake detection. Social media attracts people to share information in open environments. Facing the newly forming technical lock-ins and the loss of local knowledge in agriculture in the era of digital transformation, the urge to re-establish a farmer-centric precision agriculture is critical. The question is whether social media like Twitter can help farmers to share their observations towards the constitution of agricultural knowledge and monitoring tools. This chapter tackles the following question: which phytosanitary information can be automatically extracted from textual contents on Twitter, and what is the quality of this information? We develop several scenarios to collect tweets, then we apply different natural language processing techniques to measure their informativeness as a source for phytosanitary monitoring.

3.1 Use cases

We focus on detecting anomalies concerning crop health events. Possible anomalies include the time of the event -e.g., too early in the year-, the place of the event or the path taken by the pest, and the intensity of the attacks. In collaboration with experts in the agricultural domain from Cap2020[1] and Arvalis[2], we collected tweets concerning the following issues as observation cases:

- **User case 1: corn borer**. The corn borer (*"pyrale du maïs"* in French) is a moth native to Europe. It bores holes into the corn plant which reduces photosynthesis and decreases the amount of water and nutrients the plant can transport to the ear. Corn borers also eat the corn ear, reducing crop yield and fully damages the ear. These moths also lay their eggs on leaves of maize plant. Their larvae weaken the plant and eventually causes loss in the yield. The challenges of this use case are the following:
  - distinguish the larvae of corn borers from the larvae of other moths;
  - track their propagation timeline.
- **User case 2: yield of cereals**. The harvesting of straw cereals represents an important part of the French agricultural surface. Unexpected extreme climate events such as continuous heavy rains could result in loss in the yield. Farmers tend to express their concerns for the crops when they estimate unavoidable damages. Such concerns of yield help to predict the prices of the products. The challenges of this use case are the following:
  - index the impacted species and zones;
  - track the occurrence timeline;
  - contextualize the signals on Twitter with other data sources.
- **User case 3: barley yellow-dwarf virus (BYDV)**. The BYDV (*jaunisse nanisante de l'orge "JNO"* in French) causes the barley yellow dwarf plant

[1] `https://www.cap2020.online/`
[2] `https://www.english.arvalisinstitutduvegetal.fr/index.jspz`

disease, and is the most widely distributed viral disease of cereals. The BYDV affects the most important species of crops, reducing their yield. The BYDV can be transmitted by aphids [10]. The challenges of this use case are the following:

  - track the various symptoms depending on the species and varieties;
  - track the activities of the pest carrier of the virus in sensible season.

- **User case 4: corvids and other emerging issues**. Corvids are species of birds that include crows and ravens. Corvidea can damage crops; for example, crows can pull the sprouts of cron plants and eat their kernels. The challenges of this use case are the following:

  - distinguish tweets about the attacks of corvids, while the damaged crops can be unknown or unmentioned in the text;
  - remove noises in the data, such as mentions of the famous Aesop's Fable *The Fox and the Crow*.

To study these use cases, we conceived the following methodology:

1. For each use case, we collect tweets with an initial set of keywords and a prior knowledge of the contexts of events such as cause, results, date, and region.

2. For use case 1 and 2, we plot the historical distribution of the collected tweets to verify whether the topic popularity corresponds to prior knowledge or documented data.

3. For use case 3 and 4, as there are many irrelevant tweets in the collection, we process the collected tweets with unsupervised algorithms: Latent Dirichlet Allocation [26] and K-Means [161] to extract concepts. We examine the concepts manually with domain experts to refine the scope of the topic and eventually remove tweets outside agricultural topics.

4. For the cases with a voluminous collection of tweets such as "corn borer", "BYDV" and "corvids", to tackle the challenge of distinguish observations

from other agricultural topics like policies or advertisements of pesticide, we extract a subset of tweets (between 500 and 3000 distinct text values) to label: whether the text is about general information or a contextualized observation. From the labeled tweets, we build a classifier for event detection.

3.2 Tweet collection

3.2.1 Twitter Search API

We use Twitter API's full-archive search endpoint [3] to collect tweets. This endpoint, granted by Twitter's Academic Research product track, allows us to retrieve public tweets from the complete archive dating back to the first tweet in March 2006. This endpoint can return up to 500 tweets per request in reverse-chronological order, and pagination tokens are provided for paging through large sets of matching tweets. The API allows us to send 300 requests per 15 minutes and for each HTTP request we can add a query of 1024 characters maximum to the body of the request message. The query is a JSON object that enable us to filter tweets on certain pre-defined fields, like keywords, period, place, language and hashtag, combined with logical operators. Due to the size limit of the query, we construct filter strings to pull tweets according to what we are interested in. We will talk about the filter string construction in the next subsection.

The query also supplies optional fields to retrieve information about authors and locations. For authors, we pull the ID, the name, the display and the profile. For locations, though not available for most of the tweets, we pull the ID, the place name, the place type, the country and the geographic coordinates.

We notice that with the Search API, accented and special characters are normalized to standard Latin characters, which can change meanings in foreign languages or return unexpected results: For example, "maï" (corn) will match both "maïs" and "mais" (but). However, it is not unusual that francophones ignore accents on Twitter. Thus, we have not removed tweets that only contain normalized query words. We save original tweets as well as re-tweets.

[3] https://developer.twitter.com/en/docs/twitter-api/tweets/search/introduction

3.2.2 Filter String Construction

We collect tweets to analyze our four use cases in this chapter. We also collect tweets about other pests and diseases that will be used in Chapter 4.

- For corn borer, we filter with "pyrale". Then we annotate the text with crop names, as we explain in previous session.

  We have collected in total 16345 tweets containing "corn borer".

- For the yield of cereals, we use all the cereals in the French Crop Usage Thesaurus [146] to collect tweets (*céreales à pailles* in French). As these words are quite frequent, we add conditions to retrieve tweets containing "récolte" [4], "moisson" or "rendement" (harvest or yield in English) and to remove tweets containing "recette" or "farine" (recipe or flour in English) to construct the final dataset of 54326 tweets between 2015 and 2020.

- For BYDV, there are few tweets if we only use the term "jaunisse nanisante de l'Orge", to have as many possible tweets as possible, we applied the following filter strings:

  "jaunisse nanisante de l'Orge", "JNO", "Cereal yellow dwarf virus OR Barley yellow dwarf virus OR Luteovirus OR BYDV", "Polerovirus", "jaunisse nanisante des céréales".

  We have 3302 tweets about "BYDV".

- For corvids, some bird names are quite frequent and ambiguous. Thus, we construct the filter string by combining bird names plus terms related to crops or damages:

  "(corbeau OR freux OR corvidé OR choucas OR corneille) (récolte OR moisson OR rendement OR dégât OR degat OR détruit OR maïs OR agric OR agro OR semis OR sémence OR sèment OR semé OR parcelle OR colza OR tournesol OR cultures OR frutier OR avoine OR féverole OR céréal OR champs)".

  We have 38903 tweets about "corvids".

[4] We use "récolt" to match the different conjugations of the verb "récolter".

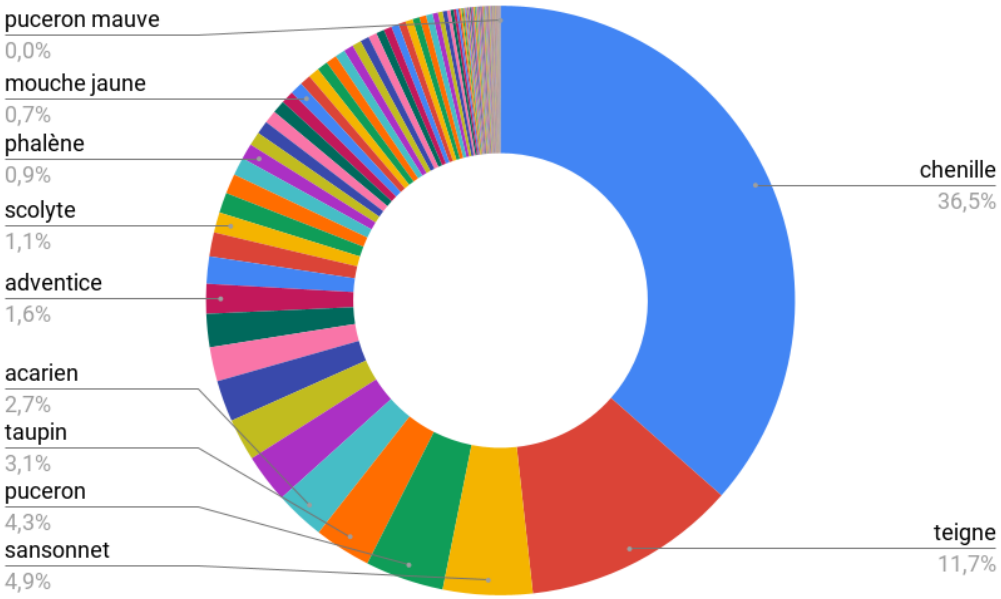


Figure 3.1: Tweets about insect pests in 2020

- To collect tweets about other pests, we crawled insect pest names from former PestObserver website [175], which tag plant health bulletins with pests and diseases. Then we use these pest names to filter tweets. Some pest names are too frequent in French expressions or too short to be unique, and bring too many irrelevant tweets. For instance, we ignore the following terms:

  “sanglier”, “mouche”, “ver”, “guêpe”, “rongeur”, “loche”, “luma”, “punaise”, “faisan”, “frelon”, “verdier”, “sansonnet”, “teigne”, “baris”, “pou”.

  We keep an exception “chenille”, to illustrate this effect in Fig.3.1, where 36.5% of the tweets about insect pests are “chenille”. We have harvested 133898 tweets mentioning pests.

- To collect tweets about other plant diseases, we concatenate the literal value of *skos:prefLabel* and *skos:altLabel* of each nodehaving type skos:Concep in “Maladies des Cultures” thesaurus [174] with “OR” logical value to build the filter string of each disease. For example, the filter string of “rouille de l’ail” is “(Rouille de l’Ail OR Rouilles de l’Ail OR Puccinia allii OR P. allii)”. Similar to the pests, we ignored the following strings:

"virus", "dépérissement", "sharka", "rouille", "pourriture", "charbon", "jaunisse", "mosaïque", "pied noir", "graisse".

We have collected in total 68683 tweets mentioning plant diseases

We upload the collected tweets to Zenodo [78].

3.3 Histogram by mention of keywords

Use case: corn borer

We plot the number of tweets by month and by year in Figure 3.2, and compare it with records of average corn borer number by trap from Arvalis (see Figure 3.3). In both figures, we can observe peaks of corn borer between May and August. There is an exception in Figure 3.2 since there are minor peaks in February, which correspond to the Paris International Agricultural Show (`https://en.salon-agriculture.com/`) when people discussed about technologies to fight corn borers. Such exception shows that tweets collected by keywords is not precise enough.

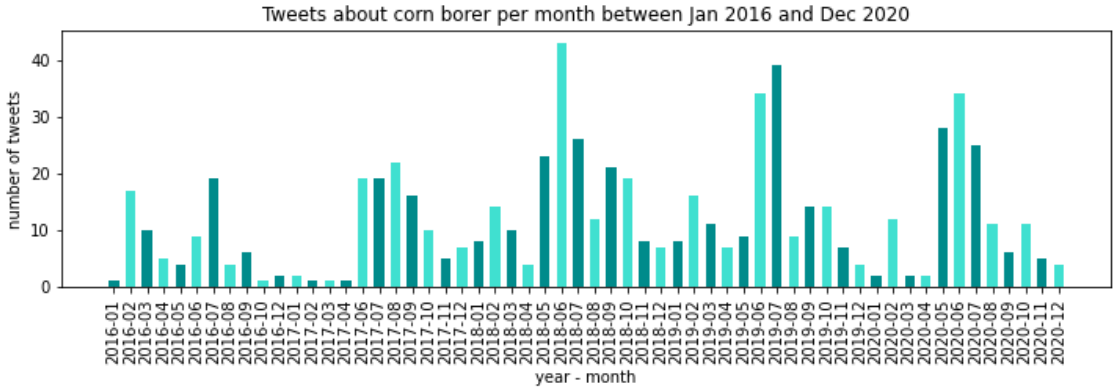


Figure 3.2: Number of tweets containing "pyrale" and "maïs" by month between 2016 and 2020.

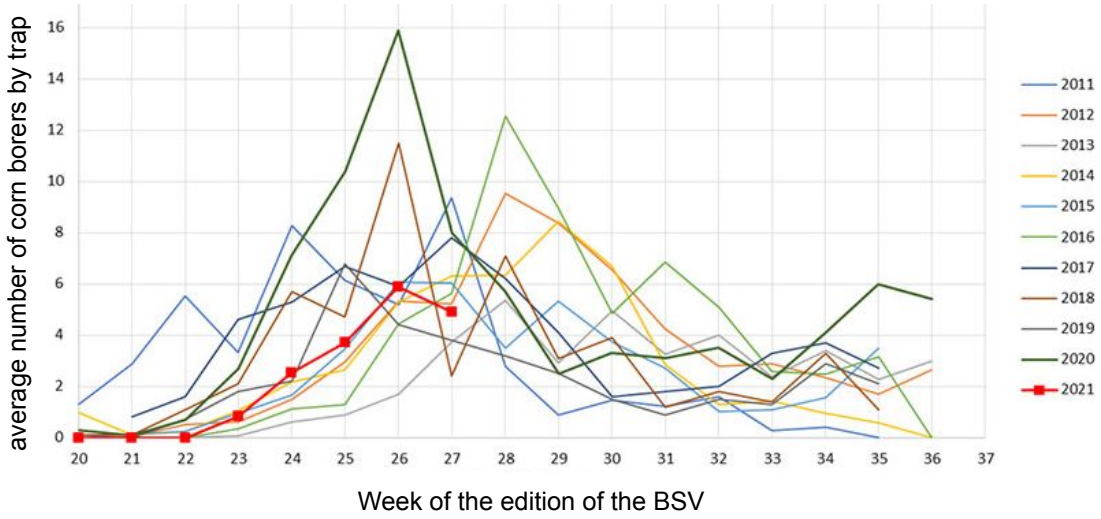


Figure 3.3: Recorded averaged corn borer number by trap from [9].

Use case: yield of cereals

Considering more and more people are engaged in broadcasting information about cereal production, we normalize the counts by using percentage of tweets mentioning cereal yields per month against the total mentions of each year and against the accumulated mentions of each month in 6 years (see Figure 3.4). Both curves show peaks between June and September each year, which correspond to the harvest season. We can also see that the peak in 2016 is higher than the other years. This abnormal popularity corresponds to the extreme yield loss in France in 2016 due to heavy rainfalls [21]. This case shows that people tend to post more tweets when bad things happen than when everything goes well, which confirms the interest of using Twitter as a source of crop health monitoring. We also plot tweets counts since this catastrophic yield containing the keywords “récolte” and “2016” in Figure 3.5. We found that this event is recalled in 2020, when people had a negative prediction for yield. We suggest that the reference of yield loss in 2016 reflects a collective memory on social media.

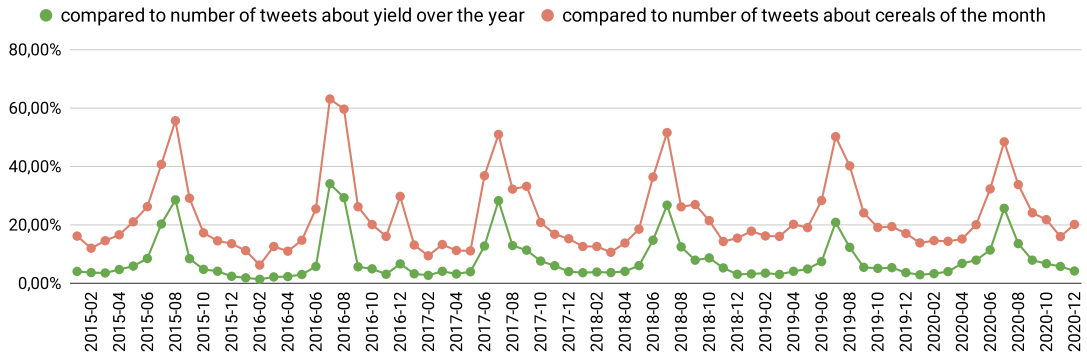


Figure 3.4: Percentage of tweets concerning cereal yield between 2015 and 2020.

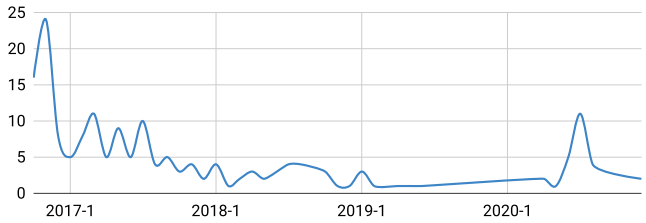


Figure 3.5: Counts of tweets mentioning yield and 2016.

3.4 Processing tweets for natural hazard detection

3.4.1 Topic detection based on Bag of Word models

As we saw in the previous section, we have collected tweets about the natural hazard of the various use cases. The goal now is to explore in detail these tweets. We can consider this task as an unspecified topic detection task [14]. Survey on topic detection [70] discussed different categories of unsupervised learning classification algorithms, including clustering techniques such as K-Means or DBSCAN, matrix factorization techniques like singular value decomposition (SVD), and probabilistic models like Latent Dirichlet Allocation (LDA) [26]. These algorithms have been created to automatically divide a collection of data into groups of similarity for browsing, in a hierarchical or partitional manner [162]. Most of the measures of similarity, such as Euclidean distance or cosine distance, can be only applied on

data points in a vectorial space [72].

Use case: barley yellow dwarf virus (BYDV)

We searched in French for "jaunisse nanisante de l'orge" or "mosaïque jaune" and its acronym "JNO". However, there are ten times as many original tweets containing "JNO" than tweets containing "jaunisse nanisante de l'orge". The reason behind this is that "JNO" is also the acronym for other things, such as "Johny's Net Online". To collect Tweets, we used all the synonyms of "jaunisse nanisante de l'orge" presented in [174]. This list also includes the keyword "BYDV", which brings also tweets in English. Therefore, we need to look into the topics in these tweets. Topics can be identified by finding the feature words that characterize tweets about the topic. At this stage, we do not know what are the topics among the tweets nor the number of topics, so we cannot use keywords to filter undesired tweets. In this sense, we isolate the irrelevant tweets with the help of a clustering method as follows:

1. Removal of stop words.
2. Calculation of the TFIDF vector for each tweet. To get a reasonable vocabulary size, we ignore terms that have a document frequency higher than 0.7 or lower than 0.01.
3. Feeding TFIDF vectors to K-Means [162], for K between 2 and 20, find the best cluster number K using elbow method [92].
4. Calculation of the TFIDF matrix for each cluster, examination of the 20 terms with the highest TFIDF scores, and manual removal of undesired clusters.
5. Repeat step 2-4 till all the clusters talk about BYDV. An example of the final state of this cleaning process is shown in Table 3.1.

We executed the same step using LDA topic modelling with the document-term matrix. Both exercises succeed to distinguish tweets in English and tweets about "Johny's Net Online" from tweets about the BYDV. We find that tweets in

English are classified to an isolated topic or cluster. We can observe “brassicole” and “hirondella” in a topic or a cluster, these are barley species that resist the BYDV. We can also see “puceron” (aphids in English) in both experiences.

Table 3.1: Top TFIDF scored words in clusters in final state of K-Means based cleaning.

cluster	top TFIDF scored words
0	année, blés, céréales, date, date semis, faire, faut, fin, jno orge, orge, précoce, pucerons, rt, variétale
1	dégâts, jno blé, orge, pucerons, rt, symptômes, virus
2	hiver, orge, orge hiver, pucerons, rt
3	année, automne, céréales, jno céréales, orge, pucerons, rt, traitement, virus, virus jno
4	année, brassicole, brassicole tolérante, brassicole tolérante jno, ceuxquifontlesessais, comportement, d’hiver, d’hiver rangs, hirondella, jno reconnue, jno reconnue brassicole, lorge, moisson, nouvelle, orge, orge brassicole, orge brassicole tolérante, orge d’hiver, orges, pucerons
5	automne, blés, hiver, jno orges, orges, orges hiver, parcelles, printemps, pucerons, rt
6	essais, faire, orge, orges, pucerons, rt, tolérantes, tolérantes jno, variétés orge, variétés tolérantes, variétés tolérantes jno
7	blé, combinaison, issue, issue combinaison, jaunisse nanisante lorge, jaunisse nanisante orge, jno jaunisse, jno jaunisse nanisante, jno maladie, jno maladie lorge, l’automne, lorge issue, l’orge issue combinaison, l’orge jno, l’orge jno maladie, maladie, maladie l’orge, maladie l’orge issue, nanisante l’orge, nanisante l’orge jno

3.4.2 Text classification based on pre-trained language models

After filtering and cleaning the collected tweets, we can be almost certain that they talk about phytosanitary issues. For plant health monitoring, there is still the need for more precision. A limit of the BoW model is that it does not represent the meaning of a word. A better feature representation technique for text classification

is a word embedding technique such as Word2Vec [58], where words from the vocabulary are mapped to N dimension vectors. Such vectors can be pre-trained on a large corpus and re-used for text classification tasks. The comparison between these vectors can be used to measure the similarity between words. Although word embedding may capture syntax and semantics of a word, it cannot keep the full meaning of a sentence [93]. Recent advancements in Bidirectional Encoder Representations from Transformers (BERT) [42] have showed important improvements in NLP, the multi-head attention mechanism seems to be promising for contextual representation. Next, we conduct supervised text classification based on a French BERT model CamemBERT [102], to verify whether CamemBERT can capture enough features of plant health observations.

Use case: corvids and other emerging issues in general

In the scenario of plant health monitoring, the incompleteness of farmers' observations on Twitter, partially resulting from the constraint on the text length, made the observation information unusable. Prior research on understanding farm yield variation [84] proposes to value them by bringing together observations from farmers and precise characterization of environmental conditions. To interconnect observation information on Twitter and other data sources, our first step is to extract tweets about observations. We define an observation as: a description of the presence of a pest or pathogens in a field in real-time. These tweets may be missing essential information, such as location, impacted crop, the developing status of the pest, damage prediction made by farmers, or suggestions of the treatment. The pest might be uncommon, as in the case of corvids, so this kind of damages are getting attention only since 2018. Thus, we can no longer filter tweets using known keywords. This observation detection is a binary classification task.

Given a small set of n labeled tweets $T = \{s_{t_1}, s_{t_2}, \ldots, s_{t_n}\}$ and a language model LM, each $s_{t_i}, s_{t_i} \in T$, is annotated with a label $o_i, o_i \in [0, 1]$ indicating whether it is of an observation. s_t can be seen as a sequence of words $s = (w_1 w_2 \ldots w_l), s \in S, T \subset S, B \subset S$, where l is the length of the sequence, w is a word in natural language. To capture the features of s, we project S to a vectorial representation X using a LM. $LM(S) \rightarrow X$ can be seen as a tok-

enizer $f(s)$ plus an encoder $g(s')$. The tokenizer contains the token-level semantics: $f(s) \rightarrow s'$ maps sequences of words $s = (w_1 w_2 \dots w_l)$ to a sequence of token $s' = (w'_1 w'_2 \dots w'_{l'})$, where w' is the index of the token in its built-in dictionary, l' is the length of this sequence of tokens. The encoder $g(s') \rightarrow x, x \in X$ transforms s' to a continuous vectorial representation x [42]. Finally, we trained a softmax classifier with X and labels of T.

We invited experts to label 1455 core borer, BYDV and corvid tweets. Then we used the pre-trained CamemBERT base model [102] to encode tweets and train the classifier. We set the max sequence length to 128 and batch size to 16. We use AdamW [100] for optimization with an initial learning rate of 2e-5. For evaluation, we plotted the precision-recall-threshold curve to find the best threshold to maximize the f1 score. To compare CamemBERT representations with BoW models, Table 3.2 shows the results of 5-fold cross validation of sigmoid classifier based on TFIDF vectors, and Table 3.3 shows the results of 5-fold cross validation of sigmoid classifier based on CamemBERT vectors. The latter is quite satisfactory. Finally, we use our classifier to predict tweets concerning natural hazards that never appeared in the training set such as wireworms ("taupin" in French, which is also a French family name). It distinguishes when "taupin" refers to a French familly name or to wireworms. For an observation such as *"Pris en flagrant délit ...M.Taupin, vous êtes en état d'arrestation #maïs #masseeds"*, even though "M.Taupin" looks like is about a person, the classifier correctly classifies it to be an observation. This means that the polysemy of "taupin" is properly handled in the contextualized embedding of the tweets, and that the classifier focus on the sense of the text beyond considering only hazard names. We present our further studies about tweet classification in Chatpter 4 and Chapter 5.

3.5 Conclusion

In this chapter, we demonstrated the potential of extracting agricultural information from Twitter by using NLP techniques. The BoW model-based data clustering proves the possibility of semi-automatically browsing topics on Twitter with explainability. The language model-based supervised tweet classification experience demonstrates that, for a given concrete NLP task, language models have the po-

Table 3.2: Classification based on TFIDF, with 5-fold cross-validation.

dataset	accuracy	precision	recall	f1
1	0.767123	0.539823	0.802632	0.645503
2	0.782759	0.566667	0.871795	0.686869
3	0.813793	0.620253	0.680556	0.649007
4	0.844291	0.702381	0.756410	0.728395
5	0.724138	0.536232	0.831461	0.651982

Table 3.3: Classification based on CamemBERT, with 5-fold cross-validation.

dataset	accuracy	precision	recall	f1
1	0.883562	0.759036	0.828947	0.792453
2	0.914384	0.857143	0.835443	0.846154
3	0.893836	0.775000	0.837838	0.805195
4	0.924399	0.913043	0.807692	0.857143
5	0.886598	0.843373	0.786517	0.813953

tential to capture their contextual information, which can reduce manual labeling work for specific information extraction. In our scenario of plant health monitoring, the extracted tweets containing observations of farmers allow us to monitor natural hazards at the field-level. Thus, we open the possibility of conducting farmer-centric research, such as analyzing and addressing the diversity of concerns and decision-making processes of different farmers. Furthermore, we can generalize our approach for the monitoring of other events on Twitter.

Chapter 4

ChouBERT: Deep Learning for Domain-specific Information Extraction

In the era of digitization, different actors in agriculture produce numerous data. Such data contains already latent historical knowledge in the domain. This knowledge enables us to study natural hazards within global or local aspects precisely and then improve the risk prevention tasks and augment the yield, which helps to tackle the challenge of a growing population and changing dietary habits. In particular, French Plants Health Bulletins (BSV, for its name in French Bulletin de Santé du Végétal) provide information about the development stages of phytosanitary risks in agricultural production [75]. However, as we reviewed in Chapter 2, the knowledge in BSV is still far from machine-comprehensive. Social media such as Twitter contain less formal than the BSV, but relevant and usually real-time hazard information. Given the nature of such publications, it is not straightforward to take advantage of their information efficiently and effectively, let alone do it automatically and relate these data to data coming from other sources such as sensors or other information systems. To handle, process and make these data searchable, it is necessary to start by classifying its textual content automatically.

Now, the idea is to automatically learn knowledge about plant health issues from BSV and apply the knowledge to improve the information extraction and

document classification from heterogeneous textual data sources. Recent advancements in Bidirectional Encoder Representations from Transformers (BERT) [42] have shown important improvements in NLP; for example, the efficiency of fine-tuned BERT models for Multi-Label tweets classification has been proved in disaster monitoring [197].

This chapter introduces our further pre-trained language model ChouBERT as an implicit knowledge base. In Section 4.1, we first examine if BERT models are capable of understanding BSV. Then, in Sections 4.2 and 4.3, we evaluate ChouBERT's "know-how" over different text mining tasks in plant health domain.

4.1 BSV classifications: understanding natural hazards

Text classification is a category of Natural Language Processing (NLP), which employs computational techniques for the purpose of learning, understanding, and producing human language content [66]. A well-known feature representation technique for text classification is Word2Vec [58] and some real-world applications of text classification are spam identification and fraud and bot detection [69]. PestObserver proposed to index BSV with crops, bioagressors and diseases based on concurrence analysis, regex and pattern matching text classification techniques [175]. However, the tags are not complete: some bulletins are only indexed with crop while the content do mention bioagressors or diseases. User-defined rules were used for relation extraction. PDF2Blocs [24] is another initiative for information retrieval from the BSVs: the script converts the PDF bulletins in french into HTML files. PDF2Blocs does not deal with the semantic of the contents. Overall, the global PestObserver approach relies on highly crowd-sourcing-dependent techniques, which makes the information extraction procedure not dynamic enough to adapt to changes in document format or contents. This section presents our early experiments before initializing collaboration with domain experts on concrete cases. Then, we explore the potential and limits of BERT considering the available data sets. More precisely, we want to answer the following questions: Will BERT be able to give an interesting classification for BSV compared to PestObe-

server? Moreover, how well can it generalize for natural hazard prediction from heterogeneous documents?

4.1.1 Experiments

Figure 4.1 illustrates the pipeline of our experiments in this section. First, we adapt the general language model to the plant health context. We further pre-train the pretrained CamemBERT model[102] and BERT-Base-Multilingual-Cased model (mBERT) [132] with the Masked Language Modelling task on raw BSV corpus in item 1.

Then we fine-tune these domain-adapted language models with two supervised classification tasks, namely *hazard classification* and *risk assessment*. The hazard classification identifies if a text talks about bioagressor, disease or both. The risk assessment task aims to test if the language model can "understand" the risk. Based on the given text, the risk classifier tells if there is upcoming damage and if yes, the hazard is a bioagressor or plant disease.

The fine-tuned model's outputs are the probabilities of all classes. We set a threshold value of 0.5 to pick up a list of possible classes to the input text as the final prediction. Finally, we evaluate the model over the validation set.

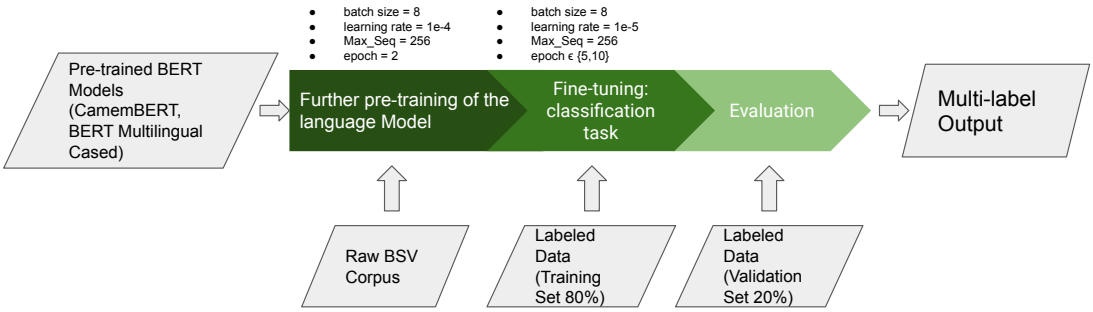


Figure 4.1: Overview of our experiments.

Data

Existing resource:

A) We downloaded BSVs [173] from the PestOberserver site. In this collection of 40828 files, there are 17286 older BSVs in XML format and 23542 OCR (Optical Character Recognition) processed BSVs in plain-text format.

B) We also obtained tags for each BSV from the PestOberserver site. There are 389 *bioagressor* tags and 279 *disease* tags, and these BSVs were annotated using text mining techniques and by domain experts. Unfortunately, only the plain-text files are annotated as *bioagressors* or *diseases*. The XML files are annotated only with crop names.

C) Tweets that we collected in Section 3.2.

Linguistic prepossessing for the text of each BSV: We removed the following from the text of each BSV:

- URLs, phone numbers, and stop words from the BSV text.
- Extra white spaces and continuous punctuation marks.
- Continuous lines that contain less than three words are rows from broken tables in the original PDF file.
- Strings like "B U L L E T I N" in vertical lines.

Dataset construction:

1. For self-supervised masked language modelling, we extracted paragraphs from XML format BSV in item A) to make the corpus.
2. For the hazard classification, we randomly split 200 cleaned BSVs into 4301 chunks containing between 5 and 256 words. We classify each chuck as *bioagressors* and *diseases* according to the tags of its corresponding BSV -see item C)-. Table 4.1 presents the numbers of labels in this dataset.

3. For the risk assessment, we manually classified 400 sentences extracted from cleaned BSVs. We classified these sentences as *bioagressor* and *disease* if the BSV says the development of a hazard reaches the threshold of danger or if it recommends applying a treatment. Table 4.2 shows the numbers of labels in this dataset.

Table 4.1: Counts of labels for hazard classification.

Bioagressor	Disease	count
0	0	2179
0	1	475
1	0	1212
1	1	435

Table 4.2: Counts of labels for risk assessment.

Bioagressor	Disease	count
0	0	193
0	1	110
1	0	87
1	1	11

Training details

All the experiments were conducted on a workstation having Intel Core i9-9900K CPU, 32GB memory, 1 single NVIDIA TITAN RTX GPU with CUDA 10.0.130, Transformers [193] and Fast-Bert [172].

For further pre-training, all the parameters of the language model are tuned on raw BSV corpus over 2 epochs as suggested in [42]. The batch size is 8. AdamW [100] is used for optimization with an initial learning rate of 1e-4.

For fine-tuning, the batch size is 8. The maximum sequence length is 256. We use AdamW [100] for optimization with an initial learning rate of 2e-5. We trained the classification model for 5 or 10 epochs and saved the one with a better F1 score.

4.1.2 Result and evaluations

To evaluate both classification tasks, we use accuracy, precision, recall, F1 score, and ROC_AUC score [68].

Table 4.3 shows results of hazard classification task with CamemeBERT. As we label the BSV blocs with the appearance of tags and the tags crawled from the PestObserver site are not completed, the pertinence of its categorization is limited. Nevertheless, we observed that the model can correct some false negative taggings from PestObserver. In other words, a phrase which mentions borer and which is not tagged as bioagressor on the PestObserver site may still be classified as bioagressor by our model. This model also shows certain generalizability when tested with tweets item C), of which the syntax is unknown to the model. As an example, consider the following text about "pyrale" (pyralid moths) from a BSV:

> *"Dans les pièges lumineux, le nombre de captures correspond à la fois aux individus mâles et femelles. Cartographie des captures des pyrales dans les pièges à phéromone dans les Pays de la Loire (Légende : vert : absence, orange : 1-4 pyrales, rouge : 5 et + pyrales)."*

For the previous example paragraph, PestObserver has no tag for it; however, our classifier predicts it to be bioagressor.

Table 4.3: prediction of the hazard (threshold=0.5)using CamemBERT model.

	Accuracy	Precision	Recall	F Score	ROC_AUC
Bioagressor	0.86	0.76	0.88	0.82	
Disease	0.90	0.69	0.88	0.77	
Weighted Average		0.74	0.88	0.80	0.91

Table 4.4: Prediction of the hazard (threshold=0.5) using mBERT.

	Accuracy	Precision	Recall	F Score	ROC_AUC
Bioagressor	0.87	0.78	0.88	0.83	
Disease	0.90	0.70	0.87	0.77	
Weighted Average		0.75	0.88	0.81	0.91

Table 4.4 shows the results of the same multi-label classification task with mBERT. The scores are slightly better than the ones produced by CamemBERT

presented in Table 4.3; however, the size of the pre-trained mBERT is bigger than CamemBERT -since it covers more than 104 languages- and it takes more time for the training.

Table 4.5 shows the risks classification task with CamemBERT model. In this experiment, manual annotation assures the training set's pertinence. We notice that the risk classifiers can detect the potential risk caused by a bioagressor or disease in the text, though the training data size is much smaller than the one of the hazard classification task. However, considering the risk level or the detection of the positive/negative sense of the phrase, the risk classifier's prediction is not that pertinent. For example, phrases like the following are still classified to having a risk of bioagressor attack even though it says there is only a few presences of bioagressor, so no action is required. These results may be improved if more data is available.

"... note l'apparition des premiers pucerons à villenauxe la petite (77) avec moins de 1 puceron par feuille. le seuil d'intervention, de 5 à 10 pucerons par feuille, n'est pas encore atteint. aucune intervention n'est justifiée."

Table 4.5: Prediction of risks (threshold=0.5) using CamemBERT model.

	Accuracy	Precision	Recall	F Score	ROC_AUC
Bioagressor	0.85	0.63	0.89	0.74	
Disease	0.83	0.72	0.59	0.65	
Weighted Average		0.68	0.73	0.65	0.91

4.1.3 Threats to validity

As we mentioned before, for the hazard classification, we simply tag the BSV blocs with the appearance of tags. Moreover, the tags crawled from PestObserver site are not completed, the pertinence of the training data is limited.

4.1.4 Conclusion

Recent advancements in BERT-based models are promising regarding natural language processing. Our objective is to classify agricultural-related documents according to the natural hazards they discuss. We have studied existing textual data in French in the plant health domain, especially the BSVs, and experimented with the mBERT and CamemBERT models. Our results show that fine-tuned BERT-based model is promising for the hazard prediction of BSV. The preliminary prediction test on tweets convinced us that BERT-based models are generalizable for representing features in the French plant health domain. We will feed our model with more pertinent data in the following works. It may also be interesting to explore alternatives such as FlauBERT [95], another BERT-based language model for French. Finally, we also plan to investigate feature-based approaches with BERT embeddings.

4.2 Tweet classification: identifying observations about natural hazards

Smart farming is an emerging concept that refers to managing and improving farming processes using modern Information and Communication Technologies (ICTs) [194]. Indeed, technology has an increasing place at tackling some of the most critical challenges in agriculture we face today [35], making a way towards the fourth agricultural revolution: agriculture 4.0 [39]. Researchers and engineers have applied a wide range of technological innovations to tackle some specific goals: they have built simulation models for climate prediction in agriculture [63], used computer vision and Artificial Intelligence to improve the production of certain types of grains [124], employ drones for soil assessment [171], implement the IoT paradigm where sensors connected to the Internet capture real-time data at the field level to monitor agricultural components such as soil, plants, animals and weather and other environmental conditions [123], and provide crowdsensing solutions to allow farmers to contribute with observations at the field level using their mobile devices [108].

Crowdsensing is a sensing paradigm that empowers ordinary people to contribute with data sensed from or generated by their sensor-enhanced mobile devices [28, 46, 54]. It introduces a new shift in the way we collect data by permitting us to acquire local knowledge through smart devices carried by people, such as smartphones, tablets, and smartwatches, among others. Furthermore, Crowdsensing allows to leverage of enhanced sensors of smartphones in a fast and economical way, in contrast to more expensive traditional methods. Driven by the increasing recognition of the importance of farming to sustain humanity and the central role of farmers in the digitisation of agriculture [90], we have witnessed the emergence of crowdsensing applications for smart farming [108].

Farmers are also increasingly present in social media such as Facebook, WhatsApp, and Twitter [169], where they voluntarily share and discuss their observations about the environment and natural events. Notably, Twitter allows farmers to freely publish short messages called "tweets" to share their observations. Taking advantage of these observations requires keeping track of relevant data sources among the noise, extracting and organizing the information they contain and sharing it with other interested users is only possible at a high human effort by manually inspecting, filtering and cleaning all data and connecting related entities and contexts. A possible heuristic way to identify the observations on Twitter could be starting from recorded issues in French plant health bulletins (BSV, for *Bulletin de Santé du Végétal* in French), filtering the tweets with known natural hazards in its context. It can be interesting to compare what farmers tweet two weeks before a pest attack gets reported in the BSV, but reading all the BSV still requires human effort. Indeed, the historical information in BSV has attracted researchers' interest. Different works have been analyzed to improve the interoperability of BSV: the VESPA project [145] annotated the BSV with region, year and crop and published the archive of BSV as linked open data; we proposed an architecture to build a knowledge graph about the crops and pests to integrate BSV and other heterogeneous forms of data in plant health domain [75]. In these previous works, the knowledge of plant health issues is yet to be extracted from the unstructured data of BSVs.

Recent applications of large-scare pre-trained language models seem promising for tackling domain-specific information extraction problems from a text in French:

Zouari [201] observed the evolution of perplexity of the models along with the further pre-training of mBERT and of CamemBERT on insurance-related text, and the authors suggest that CamemBERT adapts to French data with less time. JuriBERT [163] compared pre-training new BERT models from scratch in different architectures and further pre-training CamemBERT on French legal text for multi-class classification. Laifa et al. [94] further pre-trained CamemBERT on French Financial dataset for extractive text summarization. BERTweetFR [62] further pre-trained CamemBERT on a 16 GB dataset of French tweets and evaluated it on sentiment classification and Named Entity Recognition (NER). In this work, we propose to build ChouBERT, a pre-trained language model that "learns" knowledge in the plant health domain from BSV and recognizes similar syntax in tweets for detecting farmers' observations in the French phytosanitary context. The following two assessments drive our work:

1. **Smart farming is the key for developing sustainable agriculture** and support food needs of increasing populations [185]. Indeed, recent developments have made possible the collection of unprecedented amounts of environmental and farming data with the goal of making agricultural processes better and more efficient, thus supporting sustainable agriculture. These data are sensed from various IoT devices and processed using Big Data and Artificial Intelligence techniques.

2. **The increasing connectivity of farmers and the emergence of online farming communities**. Farmers are quickly adopting technology. They are more than ever present in social media such as Facebook, WhatsApp, and Twitter. We are witnessing the emergence of online farming communities [169]. They use such platforms to report their observations, discuss, collaboratively propose ideas, and find solutions to existing problems in online groups.

Following these two assessments, our goal is to explore the emerging application of smart farming observations via social networks -particularly Twitter- and propose an approach for tweet classification. We aim to answer the following research questions:

RQ1. how pre-trained language models (LM) can assist in the exploration of tweet-based crowd observations?

RQ2. how to further pre-train general LMs for domain-specific text classification?

4.2.1 Related work

Crowdsensing applications on social network

Farmers write down information for their for internal management [7]. The format of their notes depends on the organizational constraints and how they proceed with the information in each holding. Similar writing particularity also exists in farmers' tweets, in which information is fragmented, rare and barely visible. We can consider the monitoring information as weak signals on Twitter. Concerning weak signal observation via social networks, Vigi4med [88] developed a process to extract information mentioning undesirable effects of baclofen[1] from French forums. This extraction process includes annotation, anonymization to handle privacy and RDF generation to structure and share the data. Results in Vigi4med prove that social network contains sufficient data for evaluating certain medical problems.

Another important subject for natural disaster monitoring via Twitter is Named Entity Recognition (NER), especially for the location of tweets. For example, the Suricate-NAT platform [16], a Twitter-based crowdsensing application for natural disaster monitoring in France, proposed a Conditional Random Field (CRF)-based tool [121] for geolocation inference.

In regard to existing works on plant health monitoring using Twitter, Welvaert et al. [189] build different keyword-based queries to retrieve tweets about the Bogong moth and the Common Koel and compared the number of tweets with regularly planned surveys to validate the queries. This approach requires human efforts for building queries with hazard names or symptoms, and presents a problem for using Twitter to detect unfamiliar biosecurity events. Shankar et al. [157]

[1] A medication to treat muscle spasticity such as spinal cord injury or multiple sclerosis

gathers tweet about 14 fungal diseases and proposes supervised tweet classification with Machine Learning and word embeddings. Their good accuracy proves the feasibility of categorizing tweets for monitoring known crop stresses. However, word embedding-based representations demand for disambiguation. This work also lacks in generalizability on unknown categories of tweets. In our work, we propose to apply domain-specific contextualized embedding to improve the generalizability of classifiers on unknown hazards.

Domain-adaptive pre-training of language model

Language models (LMs) in the BERT family all use Byte Pair Encoding (BPE) or extended BPE algorithms to build their vocabularies [73, 95, 102, 195]. It relies on subword units to encode rare words without introducing unknown tokens. The vocabulary distribution in the corpus for pre-training LMs decides the input sequence of the transformer and might impact downstream tasks. We can classify the pre-training of LMs into two mainstream strategies:

- the further pre-train the weights of an existing model on domain specific corpus without touching its vocabulary and tokenizer like BioBERT [97] and [94],

- or pre-train a new LM on domain specific corpus and a tokenizer from scratch like SciBERT [20] and JuriBERT [163].

JuriBERT infers that further pre-training an existing LM can be a significant advantage compared to randomly initialised weights. At the same time, BiomedBERT [61] shows that, for biomedicine, with a large corpus, pre-training from scratch with in-domain vocabulary confers advantages over further pre-training general-domain or mixed-domain pre-trained LMs. BiomedBERT generated the vocabulary and conducted pre-training using PubMed5 abstracts.

Between the two previous approaches, ExBERT [190] proposes to add domain-specific tokens as an extension vocabulary and to reuse the weights of the pre-trained BERT model. The pre-training of ExBERT on 17 GB of domain-specific articles impacts only the embedding of the extension vocabulary and small adapter modules added to each transformer layer, which reduces the required computation resource. Furthermore, ExBERT enables the cooperation between the extension module embeddings and pre-trained BERT embeddings via applying a weighted

combination mechanism. ExBERT has been proven efficient for improving the performance of adapting a pre-training model for the target domain on named entity recognition (NER) and relation extraction (RE) tasks. However, the performance of ExBERT for text classification task shall be investigated with more experiments, because we want our classifier to pay attention to not only entire words but also the information in the morphemes so that the classifier is more generalizable to emerging issue detection on Twitter.

4.2.2 Problem Formulation

We define an individual's observation as a description of the presence of pests in a field in real time. Such observations are valuable information for discovering trends in the evolution of pest attacks and the emergence of new pests and alerting interested users. However, unlike domain-specific reporting applications, which frame observations in a predefined way, observations on Twitter are documented in free text or images. These observations may also exist among irrelevant tweets. Furthermore, these tweets may be missing essential information, such as precise location, impacted crop, the current developing status of a pest, the individual's prediction of upcoming damage, and suggestions for treatment. The knowledge that helps to recognize farmers' observations can be found in:

- vocabularies of French crop usage [146] and of plant diseases [174] as formal knowledge;
- French Plant Health Bulletins (BSV) as semi-structured domain knowledge;
- pre-trained LMs as latent knowledge representation of French language;
- tweets labeled by domain experts, containing tactical knowledge;
- unlabeled tweets concerning crops or plant health issues, as a corpus of the syntax of tweets.

Our work aims to extract individuals' observations about pests from Twitter. Our work consists of the following steps:

1. Data preparation: we collected tweets using keywords from prior knowledge, including domain experts' suggestions and existing semantic resources. Then, we invited domain experts to label a set of tweets about known issues so that we include the experts' interest in the objective of supervised learning

2. Representation learning: adjust the weights of pre-trained encoders using Masked Language Model (MLM) [73] with BSV and raw tweets to integrate the knowledge about plant health and the writing style of tweets in order to better project features of tweets in vectorial space.

3. Tweet classification: to distinguish observations from general information, advertisements, tweets of institutional accounts, and other noise.

In the classification task, we only consider the textual content of tweets, so the identification of observation-like tweets can be seen as a supervised text classification problem. Given a small set of n labeled tweets $T = \{s_{t_1}, s_{t_2}, \ldots, s_{t_n}\}$ and a language model LM. Each $s_{t_i}, s_{t_i} \in T$, is annotated with a label $o_i, o_i \in 0, 1$ indicating whether it is an observation -in contrast to general information-. Both s_t and s_b can be seen as a sequence of words $s = (w_1 w_2 \ldots w_l), s \in S, T \subset S, B \subset S$, where l is the length of the sequence, w is a word in natural language. To predict if a tweet is an observation, we propose to project S to a vectorial representation X using a LM. $LM(S) \rightarrow X$ can be seen as a tokenizer $f(s)$ plus an encoder $g(s')$. The tokenizer contains the token-level semantics: $f(s) \rightarrow s'$ maps sequences of words $s = (w_1 w_2 \ldots w_l)$ to a sequence of token $s' = (w'_1 w'_2 \ldots w'_{l'})$, where w' is the index of the token in its built-in dictionary, l' is the length of this sequence of tokens. The encoder $g(s') \rightarrow x, x \in X$ transforms s' to a continuous vectorial representation x [73]. Finally, we train a classifier $\sigma(x) \rightarrow o_{pred}$, where o_{pred} is the predicted probability of a tweet being an observation. We suggest that x encrypts not only word-level semantics but also contextual information in the domain of interest. Thus, to achieve a good representation, we study how to further pre-train the encoder with a domain-specific corpus, like the BSV B or a set of m unlabeled tweets $T_{unlabeled} = \{s_{t_1}, s_{t_2}, \ldots, s_{t_m}\}, m >> n$. We evaluate how well the encoder $g(s')$ captures the features of T by comparing predicted labels O_{pred} and the real labels O.

4.2.3 Experiments

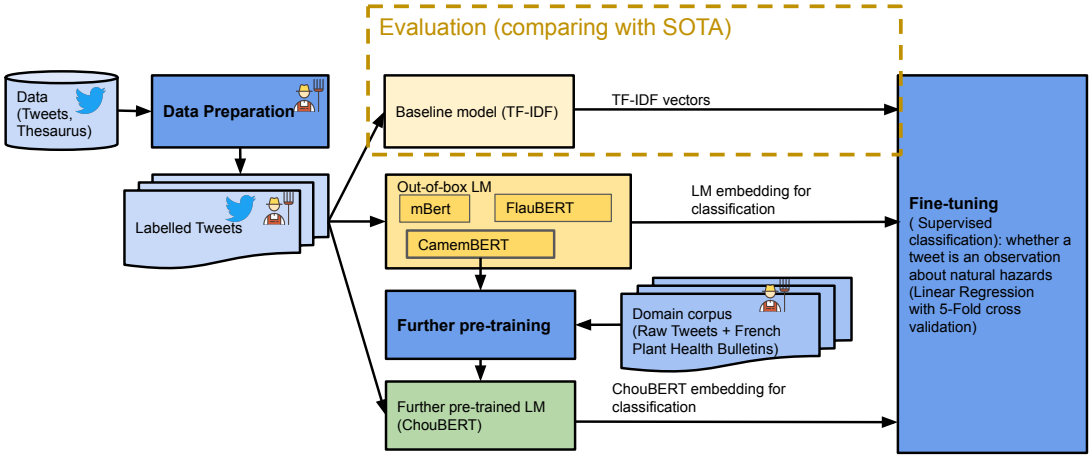


Figure 4.2: Overview of our approach.

Data preparation

The labeled set. We worked in collaboration with plant health researchers to label tweets concerning observations about three kinds o natural hazards:

- *Pyrale du Maïs* —corn borer in English— [12], representing observations about insects. On Twitter, users observe activities throughout their lifecycles, such as laying egging, the larvae attacking the corncobs, and the moth flying in the field. Users also exchange strategies to protect their cultures according to the development stage of corn borers. Depending on the context, we consider tweets discussing the current fight against corn borer as a pest observation.

- *Taupin* —wireworms in English—, is an insect pest that causes incurable damage to potatoes and cereal crops [13]. The challenge to identifying wireworm attacks is that the word "taupin" is polysemous and frequent in French.

- *Carpocapse* —coding moth in English—, similar to the corn borders, is a member of the Lepidopteran family. Its larvae damage a wide range of fruits,

including apples, pears, plums, apricots and chestnuts. This case suggests the observation of different plants and different symptoms.

- *Jaunisse Nanisante de l'Orge (JNO)* —barley yellow dwarf virus in English— [10], representing an observation about plant disease. It is a disease that can be transmitted by aphids [10]. Observations about this disease may be about the various symptoms of the disease on the crop depending on the species and the varieties, as well as the activities of the pest carrier of the virus.

- *Corvidae* —corvids in English— [11], representing observation about damages caused by birds. The dry soil in 2020 has favoured the attacks of corvids, which caused a huge loss in corn production compared to previous years, while there is no proof for the correlation between the population of corvids and the damage [8]. This case suggests the need to consider both the pest and environmental conditions. The case of corvids also represents unusual pest attacks where the cause is unknown in the early stage.

We collect tweets for each of these hazards using all the synonyms in existing semantic resources as we describe in Section 3.2. We invited domain experts to label tweets according to their judgement of their pertinence. Table 4.3 shows the composition of our labeled set. Depending on the occurrences and the impacts of each hazard, the collection periods vary for each case; for example, there were very few tweets mentioning corvids before 2015, and the very first tweet labeled as an observation of damages by corvids was sent on 27 August 2014. We collect tweets for at least two years.

We use the tweets about corn borers, corvids and barley yellow dwarf virus (JNO) to construct the training set. Of these 1358 tweets, 396 are labeled as observation (positive case). To evaluate the generalizability of our classifier on unseen hazards, we use the tweets about cording moths as supplementary training data and tweets about wireworms as supplementary test data. We chose wireworm because the word *taupin* is polysemous in French, these tweets contain many unseen noises.

hazard	French hazard name	period	total	num. of observation
corn borer 	Pyrale du Maïs	2019.1 - 2020.12	266	56
JNO 	Jaunisse Nanisante de l'Orge	2016.1 - 2020.9	625	229
corvids 	Corvidae	2009.8 - 2020.12	467	111
coding moth 	Carpocapse	2009.11 - 2021.9	362	49
wireworm 	taupin	2010.3 - 2021.9	394	33

Sources of the images:
Corn borer: https://commons.wikimedia.org/wiki/File:Ostrinia_nubilalis_(European_corn_borer),_Arnhem,_the_Netherlands.jpg
JNO: https://commons.wikimedia.org/wiki/File:Barley_Yellow_Dwarf_Virus_in_wheat.jpg
Corvids: https://fr.wikipedia.org/wiki/Corneille_(oiseau)#/media/Fichier:Kr%C3%A4he_65(loz).JPG
Coding moth: https://commons.wikimedia.org/wiki/File:2006-10-21.02._Larve_Apfelwickler.jpg
Wireworm: https://commons.wikimedia.org/wiki/File:Agriotes_lineatus.jpg

Figure 4.3: Composition of the labeled set.

The BSV We downloaded BSVs [173] from the open platform for French public data[2]. In this collection of 40828 files, there are 17286 *avertissements agricoles* (former BSVs until September 2009) in XML, and 23542 Optical Character Recognition (OCR) processed BSVs in plain-text. We first convert the XML files to plain text by extracting the heading and paragraphs. Then, we clean BSV text by removing extra white spaces, continuously repeated punctuation marks, phone numbers and strings like "B U L L E T I N " which come from vertical lines. We keep the URLs because they may contain the title of an event to which a tweet refers. The BSV contain historical reports about observed natural hazards in its region and recommendations to prevent or control the risks.

The keyword list To teach the LM the characteristic of tweets, we collect tweets containing terms in a list of 669 keyword concepts in the plant health domain between January 2015 and September 2021. We use insect pest names

[2]data.gov.fr

Table 4.6: Hyperparameters for further pre-training and fine-tuning for classification.

hyperparameters	pre-training	Fine-tuning
Batch size per GPU	[4, 8, 16]	8
Learning rate	1e-4	2e-5
Max sequence length	256	[80, 128]
Epochs	[1, 2, 3, 4, 8, 16, 32]	[4, 10]
Schedule type	warmup_cosine	warmup_cosine
Optimizer type	AdamW [100]	AdamW
Warm-up steps	-	300

in former PestObserver website [175], and the literal value of *skos:prefLabel* and *skos:altLabel* of all nodes having type skos:Concept in FrenchCropUsage thesaurus [146] and in the plant diseases thesaurus [174] to construct the list.

Experimental setup

We conduct all experiments on a workstation having Intel Core i9-9900K CPU, 32 GB memory, and one single NVIDIA GeForce RTX 3090 GPU with CUDA 10.0.130. We download the LM from Transformers [193] and use Fast-Bert [172] wrapper for further pre-training and fine-tuning. We set the hyperparameters based on the recommendation of BERT [73] and the configuration of our workstation. We do not do a grid search for all hyperparameters on all the models for simplicity.

For the further pre-training, we use implementation *CamembertForMaskedLM* in the transformer package[3]. We test different recipes to construct different corpus with the BSVs and tweets. We evaluate the further pre-trained LMs on the classification task.

Classification setup Due to the small size of our labeled data, we perform 5-fold cross-validation [139] with keeping the same separation for our labeled set to fine-tune all the pre-trained LM. Figure 4.4 shows the number of positive and negative labels in each fold of training/validation set.

[3] `https://huggingface.co/transformers/v3.0.2/`

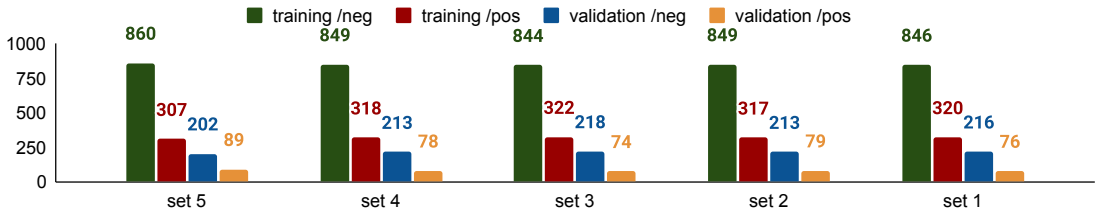


Figure 4.4: Label distribution in each fold of training/validation set.

We use the following implementations in the transformer package: *BertForSequenceClassification*, *CamemBertForSequenceClassification* and *FlauBertForSequenceClassifiaction* as classifiers, each of which is a linear layer on top of the pooled output of the LM.

Baseline model To align with the linear classifiers in the transformer package and to compare with the contextualized representations, we choose to fit the term frequency-inverse document frequency (TFIDF) vector of each tweet on linear regression classifier in sklearn package [126] to predict the probability of each class for our baseline model. To build TFIDF feature vectors, we tokenize the tweets with or without stemming and lemmatizing, then extract all the unigrams, bigrams and trigrams, and search minimum document frequency (min-df) in $[0.005, 0.003, 0.002, 0.001]$. We find that the TFIDF vectors with stemmed tokens and min-df at 0.001 give the best average precision scores on the classification task. We illustrate the scores and number of features of each fold in Table 4.7.

Table 4.7: Average precision score of baseline model.

	average_precision_score	num_feature
1	0.696615	4529
2	0.795877	4660
3	0.660066	4435
4	0.744931	4472
5	0.788443	4575
avg	0.737186	

Performance Indicator As presented above, in our labeled set, there are fewer tweets about observations (positive) than non-observation (negative), and that the positives are more important, we draw the Precision-Recall (PR) curve to evaluate each of the classifiers trained on the 5 folds of imbalanced data [152]. To have a general measure of performance, irrespective of any particular threshold, we use the average precision score in sklearn package [126], which estimates the area under the Precision-recall curve (AUCPR) [29] as the weighted mean of precision achieved at each threshold, with the increase in recall from the previous threshold used as the weight. In the following, we evaluate the models with the average of the 5 average precision scores in Table 4.9 and Table 4.10.

Results and evaluation

The tokenizers In this work, we have extracted 230 MB of text from BSVs and 20 MB of tweets to construct our corpus, which is relatively small compared to JuriBERT (4 GB for task-specific model), BiomedBERT (21 GB) or ExBERT (17 GB). Thus, we decide to further pre-train existing French or multilingual LM on our corpus, reusing the native vocabularies and tokenizers. Now the question is which LM to choose. Given the importance of the LM vocabulary to our classification task, we study the representation of our interested words produced by the following LMs: CamemBERT (camembert-base and camembert-large) [102], FlauBERT (flaubert-base-uncased and flaubert-large-cased) [95], and mBERT (bert-base-multilingual-uncased) [73]. Our initial assumption is that having more domain keywords in the vocabulary of the LM helps to produce better representations of domain text for classification.

We reuse the 669 concepts we used to collect tweets in Section 4.2.3. As words repeat in these concepts -e.g., “aleurode” (Whiteflies) in “aleurode des citrus” and “aleurode du tabac”-, we make a list of unique 599 words from the concepts. We also make a list of 549 lemmas from the words with spacy[4], so that we can compare the tokenization without considering conjugated forms. We then processed these three lists to tokenizer with the hugging face[5] implementation of each LM. Table 4.8 shows the number of broken terms. Compared to CamemBERT models,

[4] `https://spacy.io/models/fr#fr_core_news_lg`
[5] `https://huggingface.co/`

FlauBERT has larger French vocabulary sizes. Thus fewer terms are split into subwords. Though the vocabulary size of mBERT is the largest among the 5, as it is trained on 104 languages, the vocabulary size for French text is indeterminate. Similarly, there is no uncased version for FlauBERT large model. We plot the length distribution of original and tokenized tweets in Figure 4.5. On average, FlauBERT models segment less and mBERT segments more. We notice that CamemBERT-large tends to split tweets into more pieces than its base version while having the same vocabulary size but larger parameter size. Overall, most tweets are tokenized into sequences of less than 80 subword units, and the maximum length is 117.

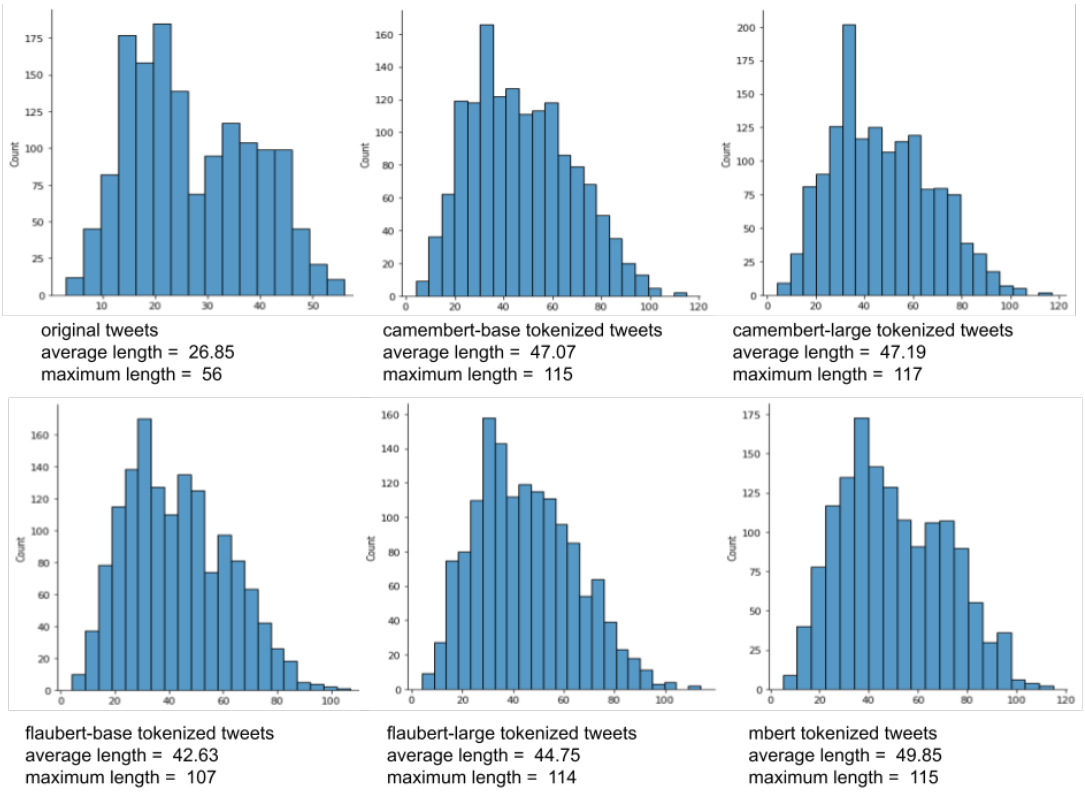


Figure 4.5: Sequence length distributions.

Finally, to evaluate how well the out-of-box LMs embed our tweets for classification, we perform the classification task described in the previous section, with max sequence lengths at 80 and 128. For each LM, we note the score with the best average precision scores in Table 4.9. All the models give better represen-

Table 4.8: Number of broken terms.

Tokenizer	vocabulary	concept	word	lemma
split by space	-	669	599	549
camembert$_{base}$	32005	617	448	393
camembert$_{large}$	32005	618	462	403
flaubert$_{base}$	67542	553	371	345
flaubert$_{large}$	68729	582	401	370
mBERT	105879	647	506	442

Table 4.9: Average precision scores of classification with out-of-box LMs.

maximum sequence length	80	128
mBERT	0.779601	0.789853
CamemBERT$_{large}$	0.863980	**0.861968**
CamemBERT$_{base}$	**0.873514**	0.855935
FlauBERT$_{base}$	0.816913	0.845478
FlauBERT$large$	0.838797	0.845027

tation for classification than the baseline model in Table 4.7, favouring contextualized embeddings. CamemBERT models outperform FlauBERT models, though CamemBERT has a smaller vocabulary and fewer parameters. There are no significant differences between the base and large models. Thus, we choose to further pre-train Camembert-base with our corpus and use the classification results of CamemBERT models as our state-of-the-art models.

The further pre-training The further pre-training is done via the Masked Language Modelling Task. Given any input sequence, 15% of the tokens are chosen randomly for prediction, of which 80% are masked, 10% are replaced with a random token, and the rest 10% remain unchanged. Then the LM is trained to predict the original token with cross-entropy [73], so it can learn the contextual information of the tokens or how the tokens are organized together. According to [142], a model trained on MLM "learns syntactic information" and "has some knowledge of semantic roles", but "cannot reason based on its world knowledge". As for our need for crowdsensing with tweets, we suggest that the LM knows how well a text looks like an observation about natural hazards, but it might not know if an observation is pertinent. As we have too few tweets labeled as observations, we let BSVs teach

Table 4.10: Average precision scores of classification with further pre-trained LMs.

maximum sequence length	80	128
ChouBERT$_{Tweets}$	0.878149	0.874741
ChouBERT$_{BSV}$	0.874490	0.865134
ChouBERT$_{BSV+Tweets}$	**0.886589**	**0.887424**
CamemBERT$_{large}$	0.863980	0.861968
CamemBERT$_{base}$	0.873514	0.855935

the LM about the context of observations and let unlabeled tweets teach the LM the language style of tweets.

We feed the out-of-box CamemBERT base model with three different groups of recipes: only BSV, only tweets or both, and we note them as ChouBERT$_{BSV}$, ChouBERT$_{Tweet}$ and ChouBERT$_{BSV+Tweets}$. We fine-tune these ChouBERT models on the classification task and note the best performance of each model in Table 4.10. We can see that all three ChouBERT models have better scores than CamemBERT models, and ChouBERT$_{BSV+Tweets}$ has the best results. It seems that CamemBERT does have the capacity to integrate the representation of tweets and of plant health from two different kinds of text for improving the downstream classification task when adequately fed.

The generalizability on unseen hazards From the previous experiments, we select the best hyperparameters (128 for maximum sequence length, 16 for batch size) for classification to study the effect of further-pre-training epochs on the generalizability of ChouBERT$_{BSV+Tweets}$ representation for detecting unseen hazards. Adding the tweets about coding moths to the previous training set/validation set of 3 hazards, we make a new set of 4 hazards for classification. We further pre-train ChouBERT$_{BSV+Tweets}$ for 0 (CamemBERT out-of-box model), 4, 8, 16, 32 epochs, train classifiers with 3-hazard set and 4-hazard set, test the classifiers on tweets about wireworm, so neither of the classifiers has seen the hazard during the training, and we plot the performance (the average of the 5 average precision scores) of each classifier in Figure 4.6.

Within the representation of each pre-trained model, the classifier trained with a 4-hazard set outperforms the one trained with a 3-hazard set, which implies

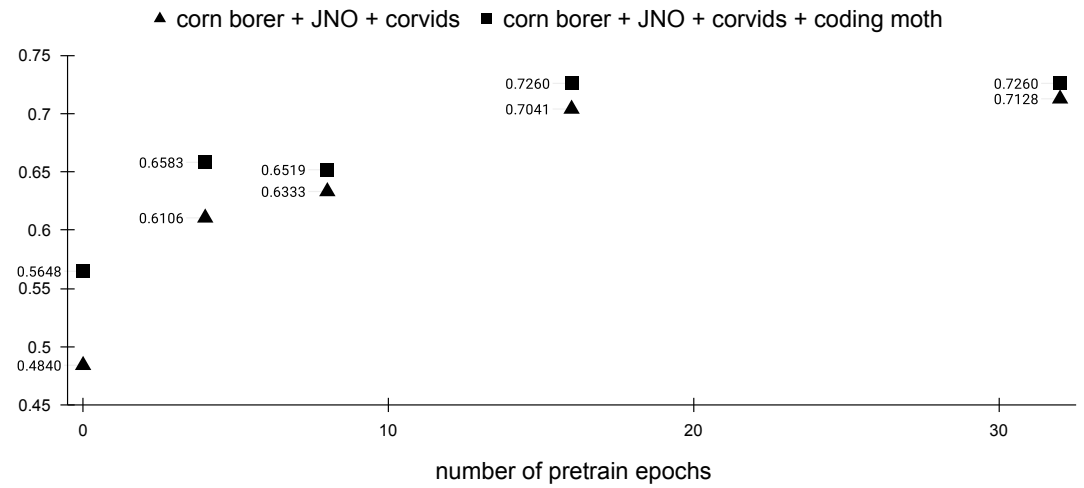


Figure 4.6: Performance of different classifiers on wireworm tweets.

that adding more training data helps improve the generalizability. With more pre-training on the text about plant health, the performance difference between the classifier trained with a 4-hazard set and the one with a 3-hazard set reduces. All the ChouBERT$_{BSV+Tweets}$ classifiers trained with a 3-hazard set significantly outperform the CamemBERT out-of-box classifier trained with a 4-hazard set. Hence, ChouBERT$_{BSV+Tweets}$ deals better with plant health-related information in tweets for classification.

4.2.4 Threats to validity

For the labeling of the tweets about observation, we did not take into account the observations concerning the absence of hazards. Based on preliminary studies about the yield of cereals in Section 3.3, we suggest that farmers tend to post their observations when bad things happen. Neither did we rank the pertinence or the completeness of the observations, which should be considered in future work. For the generalizability on unseen hazards, as we discussed in Section 4.2.3, we evaluate with labeled tweets about wireworns, which is a polysemous word in French, so the differences between in-domain tweets and the noises are significant, comparing to those in tweets containing taxonomic names. It is worth investigating to reproduce the classification task on tweets containing different unseen hazards, especially rare

hazard names out of the tokenizer's vocabulary. For the tempo-spatiality, in this study we consider all the observation tweets should be produced in real-time, if we cannot decide the temporality of a tweet, we label it as non-observation. Most of the tweets are not localized, so they can come from any francophone country.

4.2.5 Conclusion and future work

We presented a method to exploit crowd observations on Twitter. We built ChouBERT by applying domain adaptive pre-training to CamemBERT on French Plant Health Bulletins and tweets to augment the contextualized embedding of tweets for the detection of observations. We highlight the generalizability of ChouBERT representation on unseen hazards for the classification task. We can generalize this approach to improve crowdsensing based on textual content of tweets by: collecting an initial set of tweets using keywords; manually labeling a small set of tweets; further pre-training language models using domain documents and tweets; and building NLP applications with the labeled set and the domain-adapted language model. For future work, we plan to evaluate our model on other NLP tasks like NER and RE to study for the integration of heterogeneous text and the building of a knowledge base; apply our method with multilingual models on multilingual texts like the emergency bulletins of Food and Agricultural organization of United Nations; and to explore other features of tweets such as the demographic diversities in texts with contextualized embeddings. At last, our experience shows that crowd observation on Twitter is not a replacement for other monitoring paradigms but a complementary source of information. The objective of Twitter-based crowdsensing is to detect weak signals rather than quantify the gravity of an issue by the frequency of mentions. It can be interesting to cross this information with other data sources.

4.3 Model transferability to other tasks: identifying pathogens in Tweets

To continue the development of an epidemiological surveillance system based on tweets, once we find out the plant health observation information, our next step will

be identifying the natural hazards and impacted crops in such text. In information extraction, this work can be done by two tasks: Named Entity Recognition (NER) and Named Entity Linking (NEL).

Named entities are phrases that contain the names of persons, organizations, locations, in our case, disease, pest or crop. The entities and the relations among them are the essential elements of formal knowledge graph construction. We can use the knowledge graph to index and integrate information from heterogeneous documents [75]. Given a text, we split it into a sequence of tokens $S = (w_1, w_2, ..., w_n)$, the goal of NER is then to identify whether a subsequence $S' = (w_k, ..., w_l), (1 \leq k \leq l \leq n)$ is an entity. Then the goal of NEL is to assign the identified phrase to a unique concept in existing knowledge graphs. NER and NEL can be formalized as token-level classification tasks, while the classification of a tweet being an observation is sentence-level.

A recent survey about limitations of IE [1] summarises the challenges to text-based event extraction from tweets as (a) the ambiguity of representation, (b) noisy data and (c) lack of training data. The text classification results of ChouBERT in Section 4.2 prove its capacity to represent plant health-related information and filter out noises among the tweets. This section addresses the ambiguity of token-level representation and the lack of training data by examining if ChouBERT-based NER improves the detection of named entities of natural hazards in tweets with small sizes of labeled data.

4.3.1 Related work

Adnan et al. [1] categorize NER technologies into rule-based, machine-learning and hybrid approaches. In this sense, we review the available resources to analyze the feasibility of these approaches towards monitoring plant health threats on Twitter.

Rule-based approaches depend on grammar rules and dictionaries to describe human languages for computers. To achieve our domain-specific classification tasks, we generalize grammar rules and dictionaries to machine-comprehensive knowledge about plant health, like a list of known diseases of an apple variety or a list of environmental conditions that favour the germination of fungal spores. An example of domain non-specific entity extraction with rule-based systems is PADI-

web [177], a French epidemiological animal health surveillance platform. PADI-web uses rules and gazetteers to extract locations and dates. Considering domain-specific knowledge graphs about plant health, FrenchCropUsage (FCU) [146] is a French thesaurus of crops organized by usage; the project VESPA [175] produces a list of disease names and a list of pest names to index French Plant Health Bulletins using hand-crafted rules. These existing dictionaries in the French plant health domain enable information retrieval by the occurrences of terms but cannot support the disambiguation in NER nor give insights into hybrid approaches.

Machine learning approaches model a language as a probability distribution over sequences of words [87]. Then the machine can learn to classify the data points based on their similarity if the modelling extracts representative features in the text. Thus we can divide machine learning-based NER approaches into two parts: (1) converting text into vectors (aka feature engineering) and (2) applying classification algorithms to the vectors. A popular supervised classification algorithm is CRF (Conditional Random Fields) [2, 184, 196]. However, supervised approaches demand a large quantity of labeled data [1]. Semi-supervised learning can improve the classifier's performance by benefiting from unlabeled data. Gang et al. [196] propose combining two semi-supervised approaches for identifying medical concepts and annotation inconsistency: using pre-trained encoders BERT or clinical-domain BioWordVec for vectorizing the text and applying adversarial learning above CRF classifiers. BioWordVec gives static embedding to each token, and BERT gives contextualized embedding to each token. The results in [196] show that the differences between BERT-based and BioWordVec-based are always much more significant than the difference between these models and their GAN versions – with GAN or not, BERT-based outperforms BioWordVec-based, which emphasises the significance of feature engineering. Thus, we decide first to investigate if such representation improves the detection of named entities of natural hazards in tweets with small sizes of labeled data.

4.3.2 Dataset for NER

Annotation

We use disease names and pest names to construct our labeled set. For each hazard, we sample up to 5 tweets from our tweet collection described in Section 3.2. To include as many different hazards in the labeled set as possible, we do not filter these tweets with any observation classifiers from the previous session. We manually annotate 1028 tweets with INCEpTION [91] and export the labeled data with IOB2 tagging, where B- prefixed tag denotes the first term of every named entity, I- prefixed tag indicates any non-initial term and the O tag means that the token is outside any target entities. As we aim to identify diseases and pests, there are five different tags for each token: B-maladie (the beginning of any disease), I-maladie, B-ravageur (the beginning of any pest), I-ravageur and O. As the CamemBERT tokeniser could break a word into several tokens (wordpieces), we tokenise all the labeled data and label the wordpieces of a word with its IOB2 label. So, for example, the disease entity "phoma du colza" with its original IOB2 tagging [BMaladie, IMaladie, IMaladie], is tokenised intro "_pho ma _du _colza", then the labels for theses wordpieces are [BMaladie, BMaladie, IMaladie, IMaladie]. All our evaluation metrics are then calculated based on the predictions for each wordpiece.

Training-validation-test split

As our ChouBERT [76] claims its advantage in classifying unseen natural hazards and evaluates such capacity with the polysemous term “taupin”, we wonder if ChouBERT representation helps to detect unseen and ambiguous hazard names. Thus, we select a list of hazard names in Table 4.11 and use all the tweets containing such terms to make a test set of 207 tweets. Then we sampled 640 tweets to make 5-fold training-validation sets for cross-validation. At last, we append the rest 181 tweets to each of the five validation sets. So in the validation set, there are seen and unseen hazards, while in the test set, there are only unseen hazards.

Table 4.11: List of hazards in the test set and their meaning(s).

Hazard	Meaning(s)
oïdium or oidium	Powdery mildew, a fungal disease
teigne	1. an insect pest, e.g. teigne du poireaux 2. Dermatophytosis, a fungal infection of human skin 3. the title of the monarch of the pre-colonial Kingdom of Baol
rouille	1. rust (fungus), plant diseases 2. rust (iron oxide) 3. a sauce in French cuisine
mosaïque or mosaique	1. Mosaic, decorative art style 2. Mosaic virus
pourriture	1. rottenness, fungal diseases 2. political corruption
taupe	1. Mole, a small fossorial mammal 2. a family name
taupin	1. wireworm, an insect pest 2. a family name 3. an undergraduate student from a French scientific preparatory class
mouche	1. flies, insect pests 2. Bateaux Mouches, excursion boats along the river Seine
tipule	tipula, a very large insect genus in the fly family Tipulidae
cousin	1. homonym to "tipule" 2. cousin, a type of familial relationship

4.3.3 Experiments setup

We use the implementation for token classification *CamemebertForTokenClassification* in the Transformers package[6]. It loads a pre-trained CamemeBERT model and adds a linear layer on top of the token representation output. In this study, we load the out-of-box CamemBERT-base model, the ChouBERT model pre-trained for 16 epochs (denoted as ChouBERT-16), and the ChouBERT model pre-trained for 32 epochs (denoted as ChouBERT-32). Then we train the linear layer to predict the probability of a token's representation matching one of the five labels. As Croce et al. [36] claims:

> ...the quality of BERT fine-tuned over less than 200 annotated instances

[6] `https://huggingface.co/docs/transformers/model_doc/camembert`

shows significant drops, especially in classification tasks involving many categories

We are curious to see if the pre-training of ChouBERT improves the NER performance when there are less than 200 annotated instances. Thus, for each group of our training-validation dataset, we sampled 16, 32, 64, 128, 256, and 512 instances from the train set to train the NER classifier and note the precision, recall and F1 score over the validation set and the test set. We set the maximum sequence length of the model to 128. To fix the total training steps, we set the batch size to (size of the data for training / 16). Thus the number of training steps in each epoch is fixed at 16. We run all our experiments with a fixed learning rate of $5e-5$ for 20 epochs.

4.3.4 Results and evaluation

We illustrate the evolution of the average weighted F1 score of the NER classification with different training data sizes in Figure 4.7. We note the best F1 scores of 15 training epochs for each combination of unlabeled size, train size, and PLM. Remarkably, both ChouBERT-16 and ChouBERT-32 outperform CamemBERT-base initially, where we train the classifiers with only 16 labeled tweets. With less than 256 labeled tweets, there is always a significant distance between the CamemBERT-base classifier and the ChouBERT classifiers. Such distance tends to reduce when there is more data, but in most experiments, ChouBERT classifiers are better on the validation and unseen hazards sets. These results being coherent to those in [76], proves that the pre-training of ChouBERT improves different downstream NLP tasks in the plant health domain.

4.3.5 Threats to validity

As huggingface's implementation [193] of NER applies an independent linear classifier on the top of each token representation, the training does not explicitly ensure the coherence between neighbour labels. Thus, it is rare but possible that a token is classified as a non-beginning entity token like I-maladie without the token before it being classified as B-maladie or I-maladie. We evaluate the classifier's

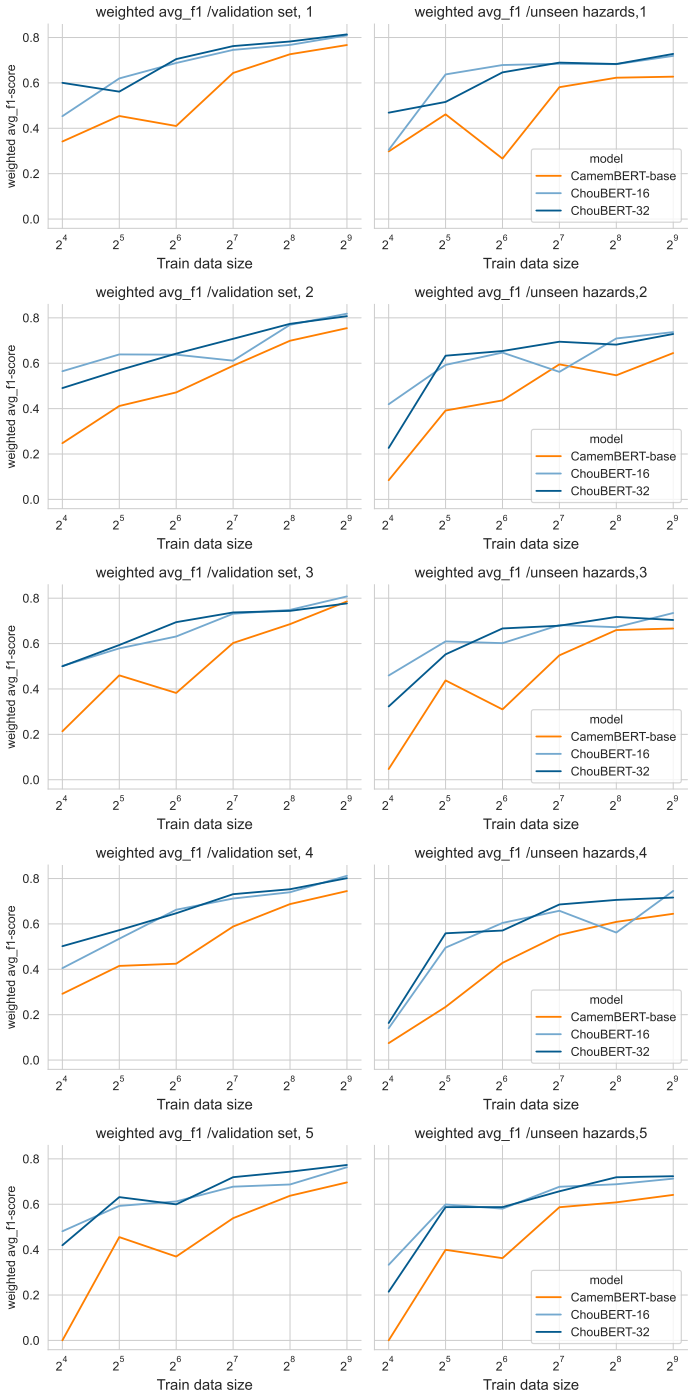


Figure 4.7: Weighted average F1 of NER on validation set and test set.

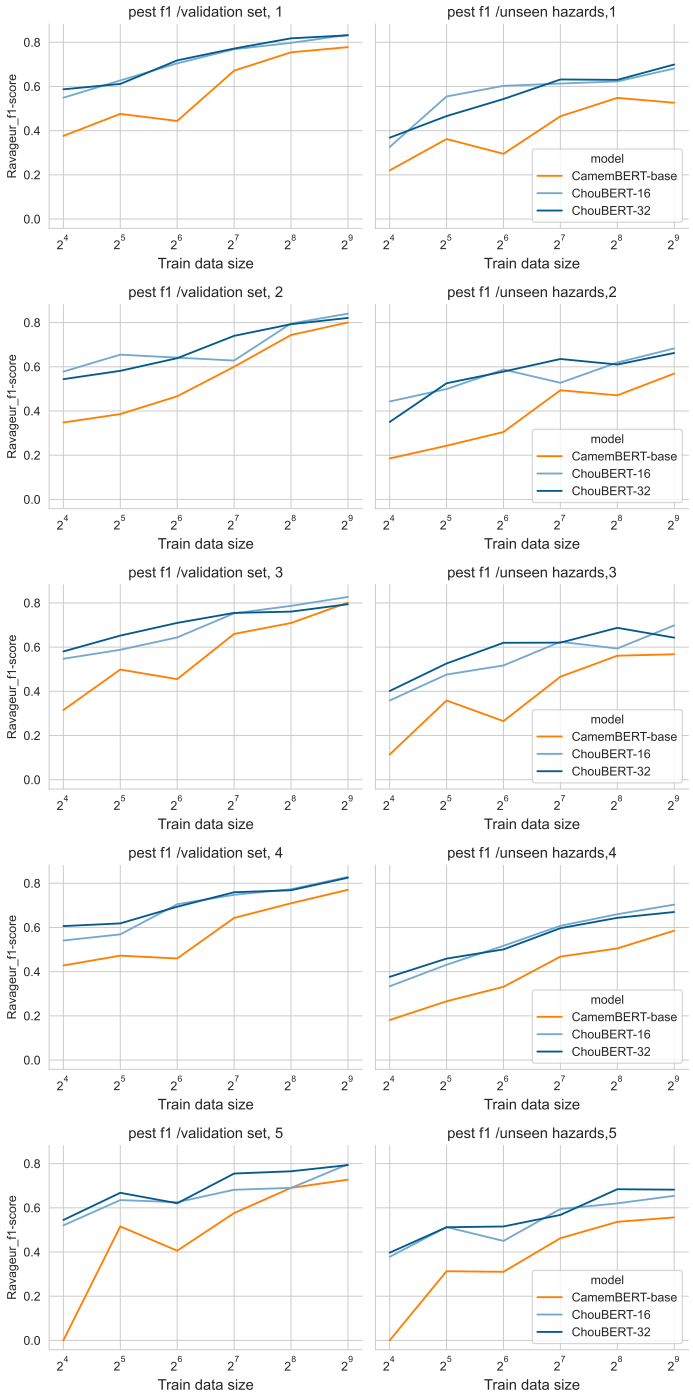


Figure 4.8: F1 of Pest NER on validation set and test set.

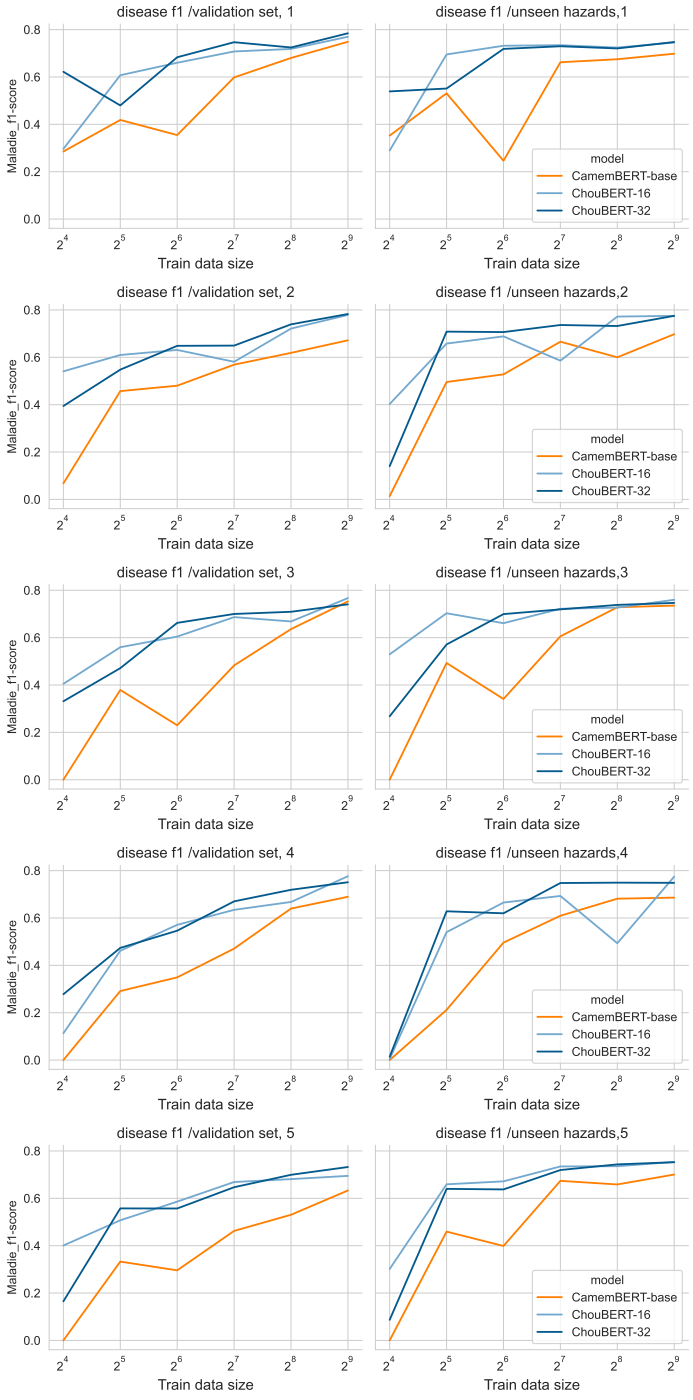


Figure 4.9: F1 of Disease NER on validation set and test set.

performance based on the prediction of each token. Nevertheless, we only combine coherent sequences when we yield the final entity annotations from the classified wordpieces.

4.4 Conclusion

In this chapter, we introduce our further pre-trained language model ChouBERT and we evaluate ChouBERT's capacity on sentence-level and token-level classification tasks in plant health domain. The output of this work is the model ChouBERT, the three ready-to-use classifiers and our labeled datasets. We have published our ChouBERT models to huggingface[7]. Our studies validate ChouBERT's capacity to detect plant health-related entities over small quantities of labeled data and its generalizability to unseen and ambiguous natural hazards. Our work opens up the fast integration of heterogeneous textual documents in the French plant health context with ChouBERT. Future directions include developing other IE tasks with ChouBERT, like entity-linking and relation extraction; annotating more fine-grained information like the symptoms and the developing stages of hazards on crops; optimizing the model with knowledge distillation; and investigating hybrid approaches with newly developed knowledge graphs.

[7] `https://huggingface.co/ChouBERT`

Chapter 5

Combining GAN-BERT setup and ChouBERT: Semi-supervised Learning for Low-resource Text Classification

PLMs suggest an objective engineering paradigm for NLP: a) pre-training language models to extract contextualized characteristics from text, followed by b)fine-tuning with task-specific objective functions [99]. We presented ChouBERT in the previous section, which has proved it a promising technology for plant health hazard detection on social media. However, we still face the lack of sufficient labeled data to validate ChouBERT's capacity on other text classification objectives, such as detecting the potential loss in yield or natural language inference (NLI). GAN-BERT extends the fine-tuning with unlabeled data in a generative adversarial setting and obtains better performance in several text classification task when the labeled set is relatively small. In this chapter, we study the combination of adversarial training and further pre-training. We will discuss 1) Does the combination improve the classification task for plant health hazard detection? 2) How does the pre-trained BERT-like model impact the training in a GAN-BERT setting?

5.1 Generative Adversarial Networks

Generative Adversarial Networks (GAN) [60] are a family of neural networks that can be commonly divided into two antagonistic parts: a generator and a discriminator, which compete during training. The generator aims to mimic real data by transforming noise, while the discriminator has to determine if a data is real or produced by the generator. The discriminator's classification results then feed the generator's training in turn. The training of GANs is known to suffer from the following failure modes: gradient vanish, mode collapse and non-convergence. Gradient vanish occurs when the discriminator cannot give enough information to improve the generator. Mode collapse occurs when the generator get stuck generating one mode. And non-convergence when the generator tends to overfit to the discriminators instead of reproducing the real data distribution.

Many variants of generative adversarial networks are proposed to improve sample generation and the stability of the training. Some of these variants are the Conditional Generative Adversarial Networks (CGANs), where the generator is conditional on one or more labels [114], and Semi-supervised GANs [153] (SS-GANs), where the discriminator is trained over its k-labeled examples plus data generated by the generator as a new label "$k + 1$" (see Figure 5.1).

5.2 GAN-BERT Architecture

GAN-BERT [36] extends the fine-tuning of BERT-like pre-trained language models (PLM) for text classification with a semi-supervised discriminator-generator setting (see Figure 5.2), introduced by [153]. Let us project all the data points in a d- dimensional hidden space, then the data vector$h \in R^d$. The generator G_{SSGAN} is a Multi Layer Perceptron (MLP) that takes a noise vector as input and tries to mimic the PLM representation of real data. The discriminator D_{SSGAN} is another MLP that gets either PLM representation of real labeled and unlabeled data $h_R = PLM(x), h_R \in R^d$, or faked representation $h_G = g(noise), h_G \in R^d$ produced by G_{SSGAN} as input, converts the input vector to inner representation $h_D \in R^d$ and performs a multi-class classification. D_{SSGAN} is trained over two objectives: 1) to correctly classify real data into K classes from labeled data

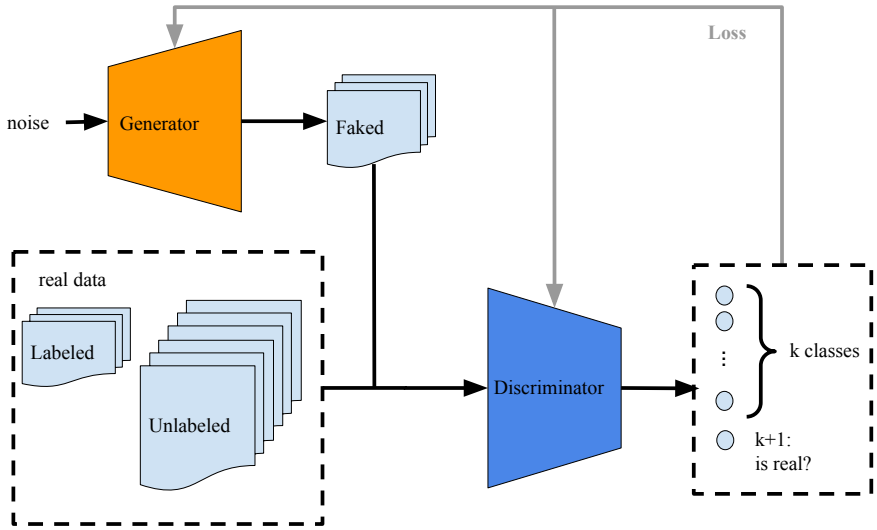


Figure 5.1: Training an SS-GAN architecture.

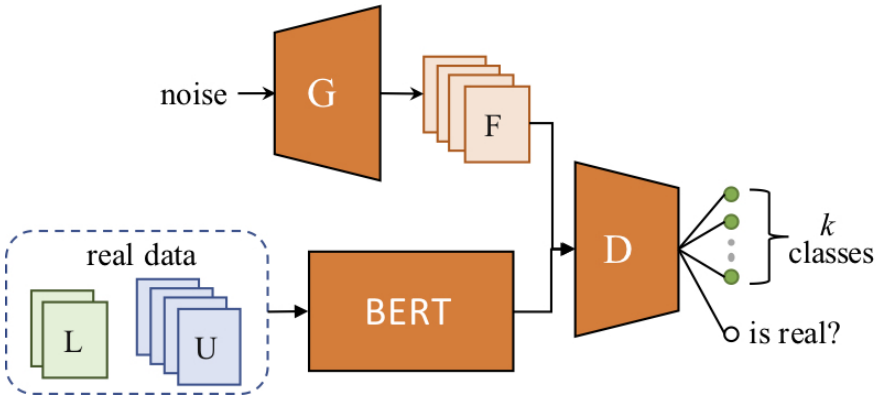


Figure 5.2: The architecture of GAN-BERT [36].

(supervised learning), 2) to distinguish generated data from real unlabeled data (unsupervised learning).

We define $p_m(\hat{y} = y|x, y \in (1, ..., k))$ as the probability given by the model m that an example x belongs to one of the k target classes, and $p_m(\hat{y} = y|x, y = k+1)$ as the probability of x being fake data. Let $\mathcal{P}_R$ and $\mathcal{P}_G$ denote the real data distribution and the generated data, respectively. The loss function for training

D_{SSGAN} becomes:

$$L_D = L_{D_{sup}} + L_{D_{unsup}} \tag{5.1}$$

$L_{D_{sup}}$ evaluate how well the real labeled data are classified:

$$L_{D_{sup}} = -\mathbb{E}_{x,y\sim\mathcal{P}_R} \log[p_m(\hat{y}) = y|x, y \in (1, ..., k)] \tag{5.2}$$

$L_{D_{unsup}}$ punishes the discriminator when it fails to recognize a fake example or when it classifies a real unlabeled example to be fake. The discriminator is free to assign any of the k target classes to the unlabeled data.

$$L_{D_{unsup}} = -\mathbb{E}_{x\sim\mathcal{P}_R} \log[1-p_m(\hat{y} = y|x, y = k+1)] - \mathbb{E}_{x\sim\mathcal{P}_G} \log[p_m(\hat{y} = y|x, y = k+1)] \tag{5.3}$$

As for the generator G_{SSGAN}, Croce et al. [36] defines the loss function as:

$$L_G = L_{G_{unsup}} + L_{G_{feat}} \tag{5.4}$$

$L_{G_{unsup}}$ penalizes G_{SSGAN} when D_{SSGAN} correctly find fake examples:

$$L_{G_{unsup}} = -\mathbb{E}_{x\sim\mathcal{P}_G} \log[1 - p_m(\hat{y} = y|x, y = k + 1)] \tag{5.5}$$

Let $f_D(x)$ denote the activation that D_{SSGAN} uses to convert the input data to its inner representation h_D. $L_{G_{feat}}$ [1] measures the statistical distance between the inner representation of real data h_{D_R} and the inner representation of generated datah_{D_G}.

$$L_{G_{unsup}} = \|\mathbb{E}_{x\sim\mathcal{P}_R} f(x) - \mathbb{E}_{x\sim\mathcal{P}_G} f(x)\|_2^2 \tag{5.6}$$

The PLM is part of the discriminator D_{SSGAN} , that is, when updating D_{SSGAN} , the weights of the PLM are also fine-tuned. And at the beginning of each training epoch, the [CLS] vector of real examples are recalculated by the updated PLM.

[1] In the pytorch implementation of the feature reg loss, the authors use: g_feat_reg = torch.mean(torch.pow(torch.mean(D_real_features, dim=0) - torch.mean(D_fake_features, dim=0), 2)) which is not exactly the same as the definition given by the paper; however, according to the author's experiments, and to our experience, there is no significant impact on the training result.

5.3 GAN-BERT Applications

By writing this chapter, the impacts of GAN-BERT have been assessed on different datasets with different PLM. The original authors of GAN-BERT have applied it to English sentence-level classification tasks, including Topic Classification, Question Classification (QC), Sentiment Analysis and Natural Language Inference (NLI) with the original BERT model [36, 73].

Later, MT-GAN-BERT [30] extends GAN-BERT to a Multi-task learning (MTL) architecture to solve several related sentence-level classification tasks simultaneously and proves to reduce overfitting. MT-GAN-BERT is assessed with English and Italian datasets, using BERT and UmBERTo[2] for sentence embedding generation, respectively. The results of MT-GAN-BERT show that GAN-BERT-based models outperform BERT-based models with 100 and 200 labeled data while training GAN-BERT with 500 labeled data, the performance worsens.

Ta et al. [165] apply GAN-BERT for paraphrase identification, propose to filter noises in the labeled set to improve the performance, and claim that in their case lower learning rate helps the model learn better. Still, a too-small learning rate makes the accuracy increase slowly. Another study that concerns noises is done by Santos et al. [155], which uses GAN-BERT with Portuguese PLMs to find hate speech in social media. This work shows that text cleaning, including removing users' mentions, links, and repeated punctuations, improves the performance of GAN-BERT-based classification. At last, the authors infer that GAN-BERT is more susceptible to noise.

In [118], the authors show that combining GAN-BERT setting with a domain-specific PLM BioBERT [97] outperforms the original GAN-BERT on a sentiment classification task for clinical trial abstracts. However, the authors do not compare the results with PLM-only classification. Neither do they provide a detailed analysis of the training? In this work, the authors have 108 labeled examples. The small number (23) of labeled samples in their test set also makes the result unconvincing, which calls for more studies to validate the combination of GAN-BERT and domain-specific PLMs.

[2]https://huggingface.co/Musixmatch/umberto-wikipedia-uncased-v1

Danielsson et al. [37] study whether and how GAN-BERT can help the classification of patients bearing implant(s) with a relatively small set of labeled electronic medical records (EMR) written in Swedish. In practice, they further pre-train a Swedish BERT model [3] to provide the [CLS] representations of 64 and 512 tokens to the discriminator of GAN-BERT and run experiments over varying training set sizes. The results by Danielsson et al. [37] show that combining GAN-BERT and a domain-specific PLM improves the classification performance in specific challenging scenarios. However, the effective zone of such scenarios is yet to be investigated.

In brief, the numerous applications of GAN-BERT witness its capacity for fine-tuning PLM on sentence-level classification tasks with low resources setting. However, none of these works above studies the correlation between the ratio of labeled/unlabeled data and the performance of GAN-BERT nor the impact of using domain-specific PLM. The lack of specifications for these hyperparameters makes the GAN-BERT setting a black box to newcomers and could lead to expensive grid search experiments for optimization [23]. In addition to that, the granularity of each classification problem divers. Therefore, it is unfair to compare the performances of GAN-BERT plus the PLMs pre-trained in different languages or domains over these tasks. In this chapter, we address these gaps by applying the GAN-BERT setting to CamemBERT, ChouBERT-16 and ChouBERT-32, and probing the different losses over varying labeled and unlabeled data size to give more insights on when and how to train GAN-BERT for domain-specific document classification.

5.4 Method

Data It is expensive and time-consuming for domain experts to annotate data, so the main challenge of detecting natural hazards in textual social media content is identifying unseen risks with low resources for training. We reuse the labeled tweets produced by Chapter 4, respectively, the tweets about corn borer, barley yellow dwarf virus (BYDV) and corvids for training and validation, and tweets about a

[3] https://github.com/Kungbib/swedish-bert-models

not only unseen but also polysemous term, "taupin" (wireworm in English) for testing the generalizability of the classifier. Since the binary cross entropy loss adopted by the discriminator of GAN-BERT favors the majority class when given unbalanced data, for the different training experiments, we sample ChouBERT's training data to 16, 32, 64, 128, 256, and 512 subsets, each subset having equal numbers of observations and non-observations. We use the same validation data and test set for all the experiments. In the validation set, there are 79 observations and 213 non-observations, in the test set, there are 58 observations and 447 non-observations.

Among the data collected by ChouBERT, there are not only a small set of labeled tweets but also many unlabeled tweets. To aliment the unsupervised learning, we have chosen from our collection (described in Section 3.2) 12308 unlabeled tweets containing common insect pest names (other than those in the labeled data) in France. We sample 0, 1024, 4096, and 8192 unlabeled data to study the effect of adding unlabeled data.

Metrics As the validation set and the test set are unbalanced, and that our interest is to find out the observations, we plot the F1 score of observation class $F1_{observation}$ and the Macro average F1 score of the whole classification.

$$F1_{macro} = (F1_{observation} + F1_{non-observation})/2 \tag{5.7}$$

Baseline model Our baseline model is a dummy and lazy classifier, if it predicts all the examples in the validation set to be non-observations, it obtains an $F1_{observation}$ of 0, an $F1_{macro}$ of 0.42 and an accuracy of 0.73; if it predicts all to be observations, it obtains an $F1_{observation}$ of 0.43, an $F1_{macro}$ of 0.21 and an accuracy of 0.27. Over the test set, an all-observation prediction results in an $F1_{observation}$ of 0, an $F1_{macro}$ of 0.47 and an accuracy of 0.89, an all-non-observation prediction returns an $F1_{observation}$ of 0.21, an $F1_{macro}$ of 0.10 and an accuracy of 0.11.

Text classification with pre-trained language model Following the work in Chapter 4, ChouBERT models are further-pre-trained CamemBERT-base model over French Plant Health Bulletins and Tweets, and that ChouBERT pre-trained

for 16 epochs (denoted as ChouBERT-16) and for 32 epochs (denoted as ChouBERT-32) are the most performant in finding observations about plant health issues. Thus, in this work, we combine GAN-BERT settings with CamemBERT, ChouBERT-16 and ChouBERT-32.

To make our state-of-the-art model, we fine-tune CamemBERT, ChouBERT-16 and ChouBERT-32 for sequence classification task over the same training/ validation / test set, by adding a linear regression layer a to the final hidden state h of [CLS] token to predict the probability of label o:

$$p(o|h) = softmax(W_a h) \tag{5.8}$$

where W_a is the parameter matrix of this linear classifier. During the training, the weights of the PLM are affected along with W_a. We develop these experiments with *CamemBertForSequenceClassification* in the transformers package[4]. To make the probability outputs from this linear regression layer comparable with the label outputs from GAN-BERT classifier, we fix a threshold to 0.5, for all the predicted probability > 0.5, we consider it as observation, else non-observation. Based on the results from Chapter 4, we fix the learning rate to 2e-5, maximum sequence length to 128 and fit the classifier for 10 epochs. We set batch size to $(training\_data\_size/8)$ to have same steps for different training data sizes.

Experimental setup For our experiments, we use GAN-BERT's latest PyTorch implementation[5], which is compatible with the transformer package. To limit the number of variables, we perform two groups of experiments, in the first group we fix batch size per GPU and epochs to 30 and train the GAN-BERT architectures over increasing labeled data sizes and unlabeled data sizes. In the second group, we fixed the training steps of each (labeled, unlabeled) pair by setting batch size to $(unlabeled\_data\_size/256)$. In the second group, we also trained the GAN network without unlabeled data – the unsupervised learning still a little from labeled data, this turn we fixed the batch size to 4 and set epochs to $(1024/train\_data\_size + \log_2(train\_data\_size))$ to approximate the number of

[4] `https://huggingface.co/transformers/v3.0.2/`
[5] `https://github.com/crux82/ganbert-pytorch`

Table 5.1: Hyperparameters for semi-supervised learning.

hyperparameters	value(s)
batch size per GPU	(training data size / 256), 32
learning rate combination (D, G)	(5e-5, 5e-5), (1e-5, 1e-5), (5e-6, 1e-6)
max sequence length	128
training data size	16, 32, 64, 128, 256, 512
unlabeled data size	0, 1024, 4096, 8192
epochs	30, 15
schedule type	warmup_cosine, None
optimizer type	AdamW [100]
warmup proportion	0.1
number of hidden layers in G	1
number of hidden layers in D	1
size of G's input noisy vectors	100
PLM	CamemBERT, ChouBERT-16, ChouBERT-32

the others with assuring that the L_{unsup} converge. Table 5.1 shows all the hyper-parameters that we tune in this work.

5.5 Results and evaluation

5.5.1 Overall metrics

We present the overall results of the fixed-steps experiments in Figure 5.3, which are the most representative and stable. We can see that comparing to PLM-only classification, PLM plus GAN-BERT setting improves the scores over the validation set and the test set of unseen hazards with 32, 64, 128, 256 training data. In Figure 5.4 we plot the performance with varying unlabeled data sizes. In both figures, we can see that the deep blues lines (ChouBERT-32) are above the yellow lines (CamemBERT), which is coherent to the results of Chapter 4 that pre-training helps improve the generalizability. The representational similarity analysis by Merchant et al. [109] shows that "fine-tuning has a much greater impact on the token representations of in-domain data" and suggests fine-tuning to be "conservative". In our experiments, we do not observe that the SSGAN setting with out-domain unlabeled data helps model generalization identifying tweets

about upcoming unseen hazards. For small training data sizes, adding unlabeled data helps improving the performance on the test set, but adding more and more unlabeled data consumes more computation resources without making significant difference. Similar phenomena can be observed in the results of the fixed batch size group in Figure 5.5, where adding more unlabeled data brings more training steps per epoch, and eventually better reduces the L_{sup} within the same training epochs. In all our experiments, with 512 labeled data, PLM-only solutions outperforms PLM+GAN-BERT setting while PLM+GAN-BERT improves the performance over the validation set and over the test set with between 32 and 256 labeled data, which corresponds to the results by breazzano et al. [30] and by Danielsson et al. [37].

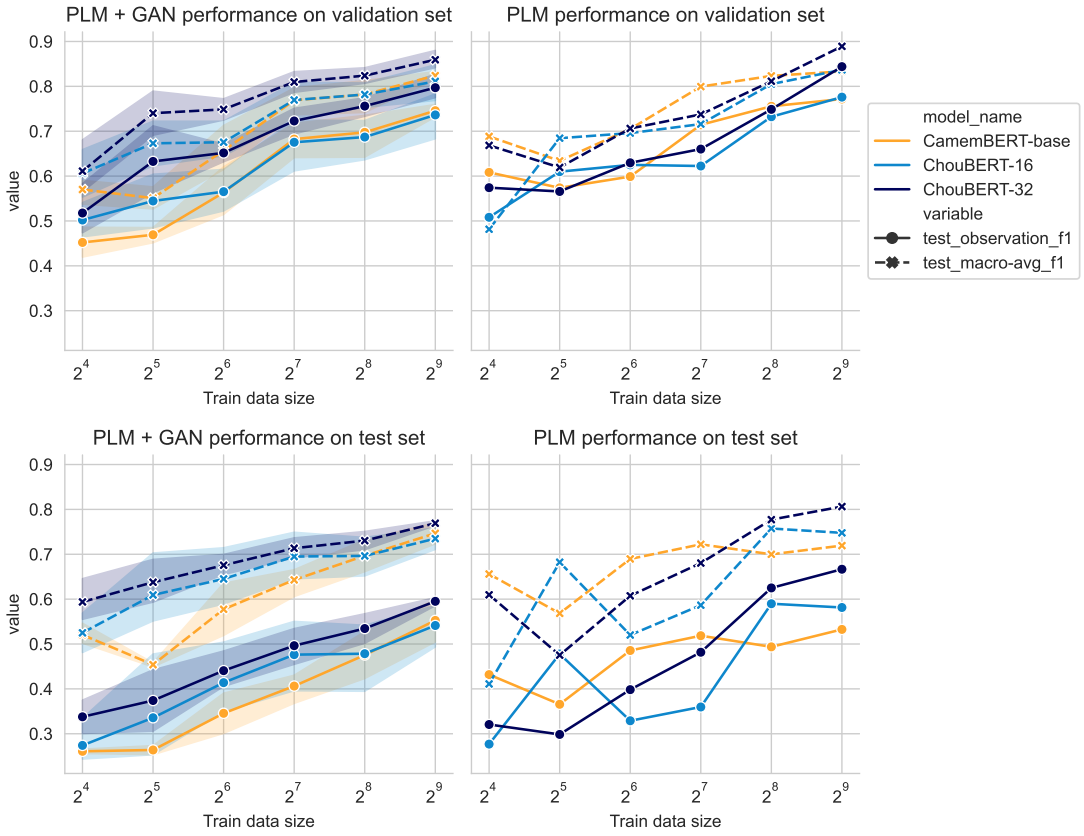


Figure 5.3: PLM + GAN-BERT vs PLM only, with same training steps over varying training data sizes.

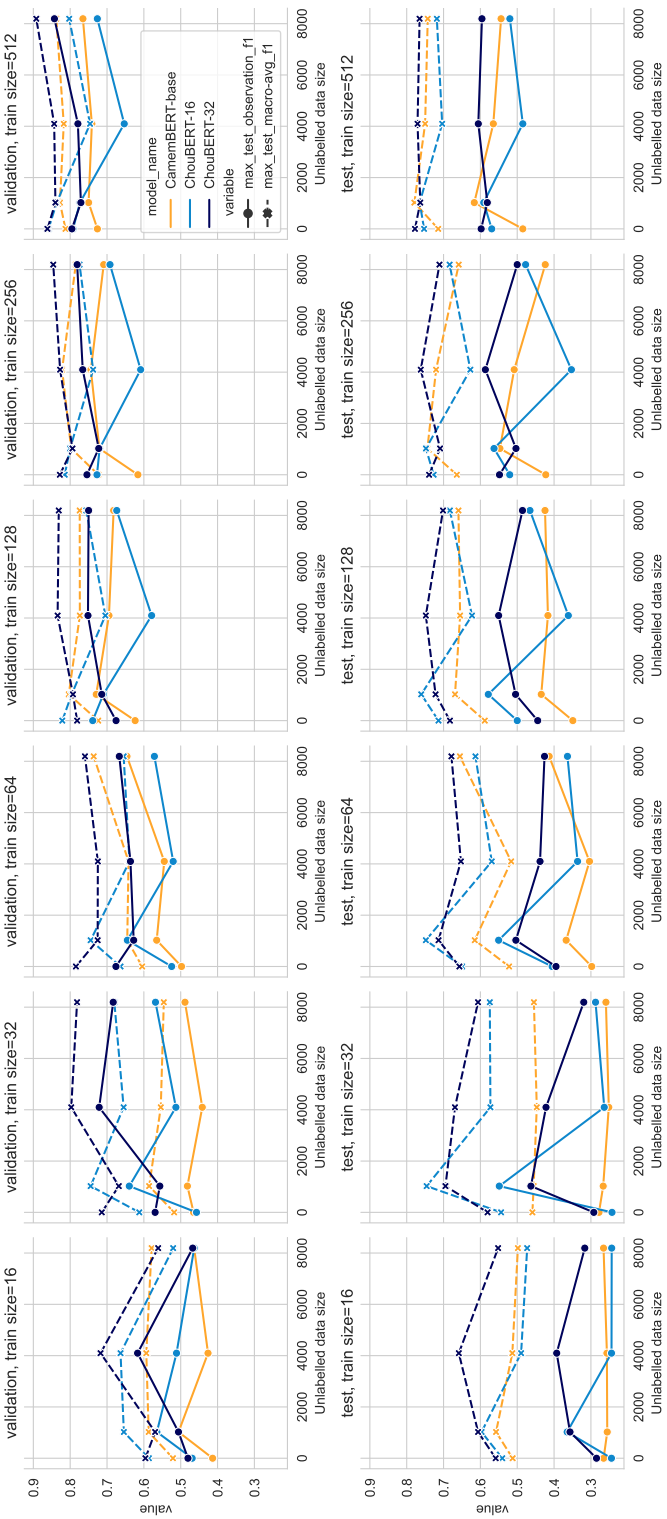


Figure 5.4: PLM + GAN-BERT performance with fixed steps, training data set sizes = 16, 32, 64, 128, 256, 512

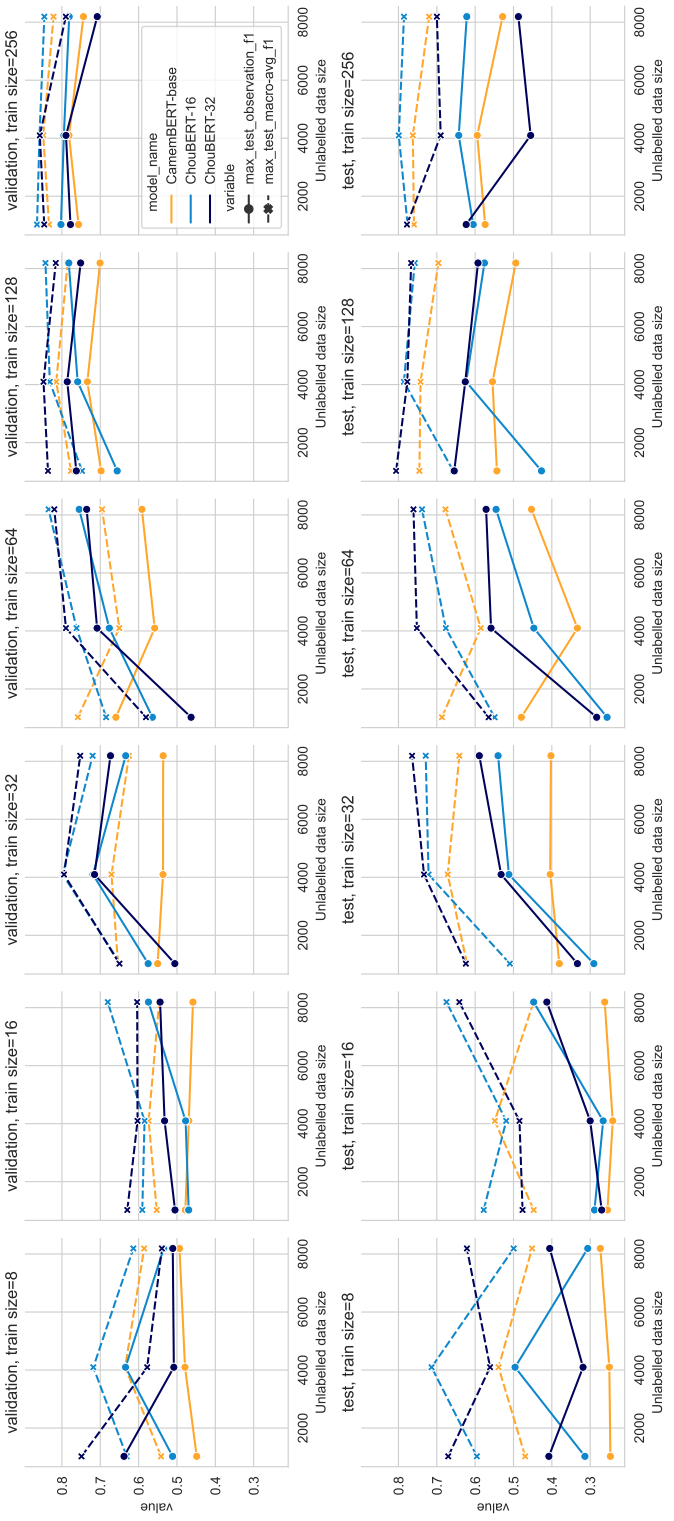


Figure 5.5: PLM + GAN-BERT performance with fixed batch size to 32, training data set sizes = 16, 32, 64, 128, 256, 512

5.5.2 The instability of GAN-BERT setting with ChouBERT models

Even though the fine-tuning of pre-trained transformer-based language models such as BERT has achieved state-of-the-art results on NLP tasks, fine-tuning is still an unstable process. Indeed, training the same model with multiple random seeds can result in different performances on a task [116]. In Figure 5.6a, we illustrate the training losses of the discriminator and the generator when given different sizes of labeled data with fixed unlabeled data size and learning rate. The trainings with ChouBERT models have more difficulties to converge than with CamemBERT. Thus, we explore the evolution of different losses and the classifiers' performance metrics on the validation set and on the test set in Figure 5.6b and Figure 5.6c, where the discriminators' losses with ChouBERT-16 take more epochs to decrease than with CamemBERT. We can observe that discriminators' losses have the same shape as L_{sup}. In particular, to present the evolution of $L_{G_{feat}}$ at the same scale as the other losses, we multiply its value by 10 to draw its line. We can see that with ChouBERT-16, the L_{sup} has more difficulties decreasing than with CamemBERT. We interpret the increase of $L_{G_{feat}}$ as that the generator tries to catch up with the fine-tuning of PLM, and the decrease of $L_{G_{feat}}$ towards its initial value as that the major changes of fine-tuning are done.

According to the authors of SSGAN [153], "in practice, L_{unsup} will only help if it is not trivial to minimize for our classifier and we thus need to train G to approximate the data distribution", which explains that while the L_{unsup} of D and G converge at the same rhythm with CamemBERT and with ChouBERT-16, the troubled decreasing of L_{sup} with ChouBERT-16 renders worse F1 scores than those with CamemBERT. For example, in the group with 16 training examples (see Figure 5.6b), the test $F1_{observation}$ scores with ChouBERT-16 are switching between 0 and 0.43, which means that the classifier predicts either all as non-observation or observation, as we describe in Section 5.4. Considering the unbalanced nature of our validation set and test set, all-observation predictions and all-non-observation predictions are two local Nash equilibria to the training of our SSGAN.

It is also remarkable that in the group with 64 training examples (see Figure 5.6c), ChouBERT-16 gives better F1 scores than CamemBERT in the early stages.

However, after the bounces of L_{sup}, though the fine-tuning helps it to decrease again, the F1 scores are not as good as before, because the effect of L_{unsup} is already gone. We can also observe this in the group with 32 training examples. More dramatically, when repeating the experiments with the same hyperparameters, the "troubled decrease" of L_{sup} does not always happen, but statistically, most of them come with ChouBERT models, especially ChouBERT-16. Our strategies against the "troubled decrease" include:

- using a smaller learning rate with more training epochs at the cost of computational resources (see [165]);
- applying a smaller learning rate to G than to D (see [65]).

Applying schedulers, and down-sampling the majority class to balance the training data – in our case, the up-sampling propose by the original code of GAN-BERT do not help. With the optimizations mentioned above, L_{sup} with CamemBERT decreases at a steady pace, and "troubled decrease" happens less often with ChouBERT models. When we look into the embeddings of [CLS] produced by the PLM, we find that there are more variance in each dimension of CamemBERT embeddings than in each dimension of ChouBERT embeddings, before and after the fine-tuning, $Var_{CamemBERT} > Var_{ChouBERT-32} > Var_{ChouBERT-16}$. Thus, ChouBERT models produce more homogeneous encodings then CamemBERT. This explains why ChouBERT embeddings are more generalizable for detecting unseen hazards: the embeddings of texts containing unseen hazards are more similar to those of seen hazards, so the downstream classifier is more familiar with these vectors. From the other side, it also means that the differences between observations and non-observations are more subtle in ChouBERT's latent space. Thus, the training of GAN plus ChouBERT needs lower learning rates to converge, while GAN plus CamemBERT is robust to converge in most configurations.

5.6 Threats to validity

In this section, we discuss the limitations of this works.

- By comparing the maximum F1 scores of each configuration during the training in the Figures 5.3, 5.4, and 5.5, we assume the performance of the classifiers over validation set and test set to be continuous and relatively stable in a period once the training converges, that is, overfitting will not cause huge drops immediately. Actually, as the validation set and the test set are imbalanced, it is easy to prove that the F1 score (the objective of our classification task) and the binary crossed entropy loss (the objective of GAN-BERT's training) are not completely aligned and may lead to sub-optimal convergence.

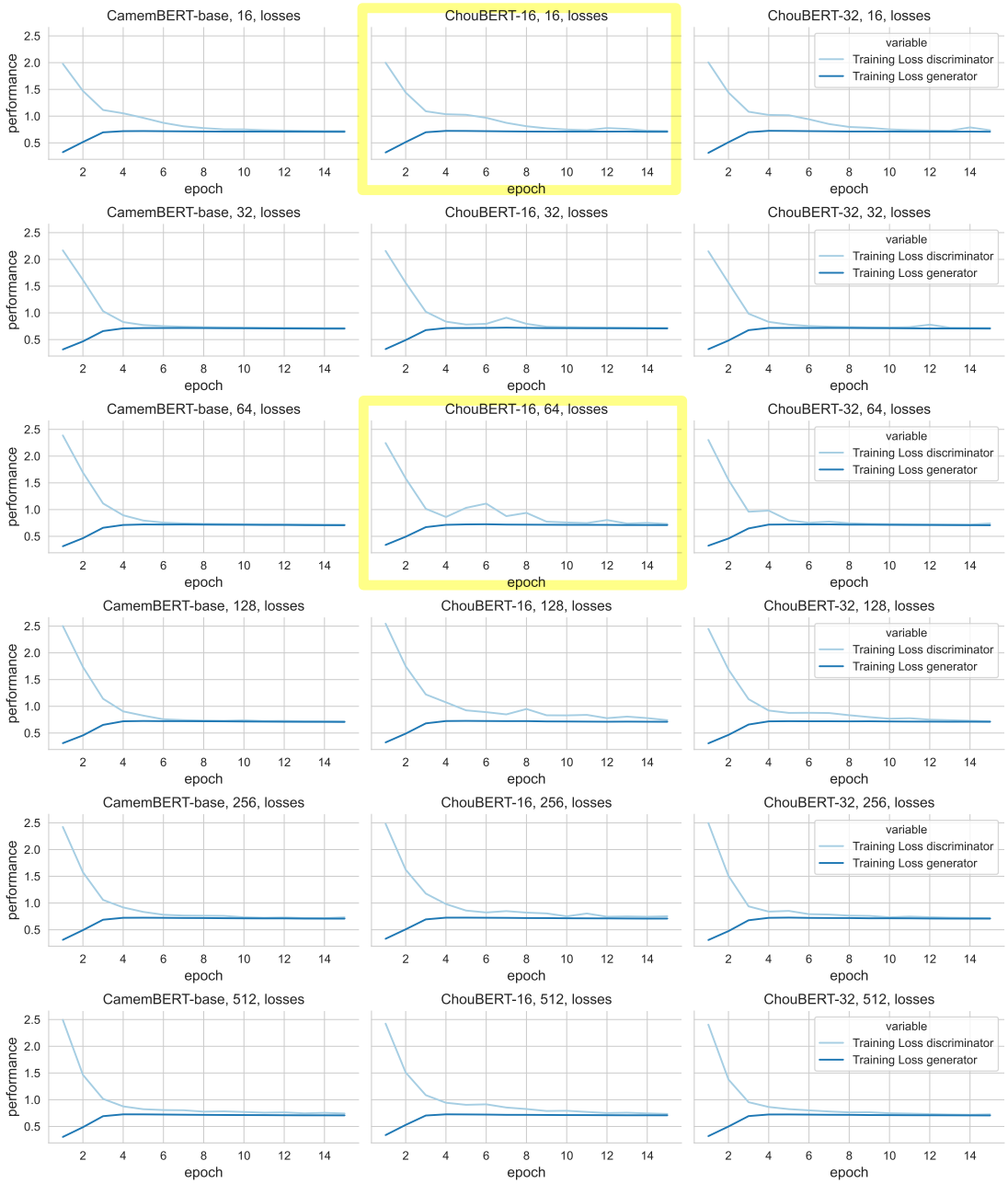
- For the PLM-only classification, we use 0.5 as a threshold to simplify the comparison with the PLM plus GAN-BERT classification. When applying the PLM-only classification to other datasets in other domains, one might need to find an optimal threshold depending on the real needs for precision or recall.

5.7 Conclusion

In this chapter, we demonstrate that combining ChouBERT models and GAN-BERT benefits from the generalizability of the domain-specific PLM to classify unseen hazards, but training such SSAN could also suffer from extra instabilities compared to using GAN-BERT with CamemBERT, a general PLM. Our experiment results validate that GAN-BERT setting improves the task of natural hazard classification when given between 32 and 256 labeled data.

Based on our experimental studies, we give our suggestions to reduce the instability: (1) The L_{sup} needs a certain minimum number of steps to decrease to zero. For a fixed batch size, adding some unlabeled data makes more training steps to go through in each epoch, consequently helping L_{sup} to decrease at a similar pace as L_{unsup}. When the number of unlabeled data is limited, using smaller batch sizes and training for more epochs helps too. (2) If the task is not too domain-specific, in other words, when the further pertained language model cannot significantly outperform the general language model in the PLM-only classification, using a general language model with GAN-BERT setting is safer. On the other hand,

if the task is highly domain-specific, apply schedulers, down-sample the majority class to balance the training data, and use smaller learning rates to train GAN-BERT with further-pre-trained language models. (3) Choose a suitable PLM. This suggestion is a question for future studies: we observe that ChouBERT-32 outperforms ChouBERT-16 in an SSGAN setting, but how to further pre-train PLMs to adapt better SSGAN setting is yet to investigate.



(a) A macroscopic view of the evolution of training losses, with fixed unlabeled size to 4096, Learning rate of Discriminator and Generator to 5e-6 and 1e-6, over 15 epochs.

Figure 5.6: Evolution of different losses and performance evolution, with fixed unlabeled size to 4096, Learning rate of Discriminator and Generator to 5e-6 and 1e-6, over 15 epochs.

(b) training data size = 16.

Figure 5.6: Evolution of different losses and performance evolution, with fixed unlabeled size to 4096, Learning rate of Discriminator and Generator to 5e-6 and 1e-6, over 15 epochs.

(c) training data size = 64.

Figure 5.6: Evolution of different losses and performance evolution, with fixed unlabeled size to 4096, Learning rate of Discriminator and Generator to 5e-6 and 1e-6, over 15 epochs.

Chapter 6

Conclusion and Perspectives

6.1 Conclusion

This thesis proposes: (1), to use Twitter as an open crowdsensing platform to acquire people's perceptions of crop health so we can include farmer participation in the reconstruction of agricultural knowledge; and (2), to serve pre-trained language models as an implicit and domain-specific knowledge base that integrates heterogeneous texts and supports information extraction from text. Boosted by the advancement of NLP and machine learning technologies, our approach facilitates information extraction from textual when semantic resources are still limited, and helps populate the explicit knowledge graphs. More importantly, our application on Twitter data involves more human contributions to the knowledge acquisition, opening up the integration of farmers' intelligence in intelligent plant health sensing paradigms.

Chapter 2 makes a landscape of existing resources for textual data integration in agriculture to evaluate the feasibility of stochastic approaches and linguistics approaches. The reviewed works concerning about existing knowledge graph in the plant health domain present approaches for data description using ontologies and RDF. Still, they do not deal with automatically mapping heterogeneous data sources to build such knowledge graphs in plant health. NLP technologies, notably PLM, seem to be promising in other domains. But how PLM works is yet to be explained. Our most essential need for information extraction in a plant health

monitoring context is to provide reusable formal knowledge to the computer and facilitate classification tasks. We propose to build PLM as an implicit knowledge base, which gives intuitions to the machine to extract information and populate explicit knowledge like knowledge graphs.

Chapter 3 records different stages in developing our proof-of-concept (POC) when initializing collaboration with agronomic experts. At the cold start stage, we do not have any concrete research topics like a specific pest and the agronomic experts cannot imagine the extractable information from Twitter nor the granularity of such information. Thus, we begin to demonstrate the power of NLP with fast-to-build, basic, explainable and unsupervised approaches, including (1) plotting counts of popular keywords by month and (2) clustering topics about plant health issues on Twitter with the BoW model. Then the experts figure out different qualitative and quantitative use cases in return. Next, we collect, clean and pre-analyze tweets for each use case and ask domain experts if such tweets satisfy their needs. Next, the experts examine and label useful tweets for them: the individuals' observations about natural hazards. Finally, we build classifiers with pre-trained language models to validate that these tweets are "identifiable" with existing technologies. Our experience shows that crowd observation on Twitter is not a replacement for other monitoring paradigms but a complementary source of information. The objective of Twitter-based crowdsensing is to detect weak signals rather than quantify the gravity of an issue by the frequency of mentions. It can be interesting to cross this information with other data sources. Our approach is easily adaptable for other NLP-driven multi-disciplinary research, such as computational social science.

Following the POC, in Chapter 4 we study the potential of pre-trained language models over plant health corpus in detail. We built ChouBERT by applying domain adaptive pre-training to CamemBERT on French Plant Health Bulletins and tweets to augment the contextualized embedding of tweets to detect observations. We highlight the generalizability of ChouBERT representation on unseen hazards for the classification task. Then our natural hazard entity detection experiments prove that the pre-training of ChouBERT can also benefit token-level NLP tasks in the plant health domain. We generalize our approach to improving crowdsensing based on the textual content of tweets as the following steps:

first, collecting an initial set of tweets using keywords; second, manually labeling a small set of tweets; third, further pre-training language models using domain documents and tweets; finally, building NLP applications with the labeled set and the domain-adapted language model. That is how we use ChouBERT to integrate the domain information contained in BSV and Tweets and convert the information into implicit "know-how" to guide the acquisition of explicit knowledge.

Chapter 5 further studies the performance of ChouBERT on the text classification to tackle the lack of labeled data in the plant health domain. Our experimental results show that combining ChouBERT and GAN-BERT can still benefit from the generalizability of ChouBERT to classify unseen hazards. However, training such GAN architectures could also suffer from extra instabilities compared to using GAN-BERT with a general PLM like CamemBERT. These instabilities open up another way of evaluating further pre-trained language models and call for future studies with other corpus and PLMs.

6.2 Perspectives

We organize the future works in two aspects: industrialization and further research directions.

For industrial usage, our ChouBERT models follow the PyTorch implementation of transformers and are ready to deploy. For example, we prototype a dashboard for monitoring the plant health on Twitter. On a local server, we pull tweets, annotate the tweets with controlled vocabularies, classify the tweets with ChouBERT's text classifier and push the classified tweets to a Google Sheet. We visualize the Google Sheet in a Google Data Studio application. The Google Sheet allows us to grant read/write roles to different users. In an active learning setting, the users propose to correct the classification in the Google Sheet and trigger a new training of the ChouBERT-based tweet classifier. Concerning ChouBERT's NER classifier, we suggest either using it directly or wrapping it into a plugin for any open-source annotator that supports active learning like INCEpTION [91].

For further research directions, first, ChouBERT opens up many exciting research topics in BERTology [142], such as probing the knowledge inside the language model [130, 140, 187], few-shot learning [56, 166], building multilingual

models to improve the classifier's performance with low-resources and unbalanced labeled data [117], creating multimodal models with knowledge graph embeddings like [17], or applying ChouBERT to other information extraction tasks like entity-linking and relation extraction. We can also investigate the model optimization, such as trying smaller architectures or applying knowledge distillation technique [154]. Second, as we model ChouBERT as a formal implicit knowledge base, the newly developed knowledge graphs (formal explicit knowledge bases), should "internalize" into language models. A direct example is to guide the information extraction with the rules in the ontology (ontology-based information extraction) [43, 192] and feed the more relevant texts to the pre-training and fine-tuning of ChouBERT. Third, applying the text mining to other kinds of texts. Though our experiments are limited to the textual contents in tweets and BSVs, our solution can take any text in French as input data. An interesting direction is reconstructing and recontextualizing the traditional knowledge in agriculture via mining proverbs and farmers' sayings. Recently published, AlemBERT [53] is a RoBERTa based language model trained on a large Early Modern French corpus (historical French from the 16th to the 18th centuries). The AgroCCol project [120] aims to analyze the modes of elaboration and transmission of ancient agronomic knowledge based on a digital corpus extended about agronomic works of the ancient times. We eagerly look forward to seeing if further pre-training AlemBERT on the corpus of AgroCCol makes a ChouBERT for traditional agricultural knowledge acquisition.

List of Tables

List of Figures

Bibliography

[1] Kiran Adnan and Rehan Akbar. "Limitations of information extraction methods and techniques for heterogeneous unstructured big data". In: *International Journal of Engineering Business Management* 11 (2019), p. 1847979019890771.

[2] Sergio Torres Aguilar and Dominique Stutzmann. "Named entity recognition for french medieval charters". In: *Workshop on Natural Language Processing for Digital Humanities*. 2021.

[3] Katie Allen et al. "A little birdie told me about agriculture: Best practices and future uses of Twitter in agricultural communications". In: *Journal of Applied Communications* 94.3 (2010), pp. 6–21.

[4] Oscar Alomar et al. "Development and testing of the media monitoring tool MedISys for the monitoring, early identification and reporting of existing and emerging plant health threats". In: *EFSA Supporting Publications* 13.12 (2016), 1118E.

[5] Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D. Manning. "Leveraging Linguistic Structure For Open Domain Information Extraction". In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Beijing, China: Association for Computational Linguistics, July 2015, pp. 344–354. DOI: 10.3115/v1/P15-1034. URL: https://aclanthology.org/P15-1034.

[6] Fatima Ardjani, Djelloul Bouchiha, and Mimoun Malki. "Ontology-Alignment Techniques: Survey and Analysis." In: *International Journal of Modern Education & Computer Science* 7.11 (2015).

[7] Mazé Armelle et al. “Entre mémoire et preuve : le rôle de l’écrit dans les exploitations agricoles”. fr. In: *Natures Sciences Sociétés* 12.1 (Jan. 2004), pp. 18–29.

[8] ARVALIS. *Comment lutter contre les dégâts de corvidés*. https://www.arvalis-infos.fr/comment-lutter-contre-les-deg-ts-de-corvides--@/view-33875-arvarticle.html. 2020. (Visited on 11/08/2021).

[9] ARVALIS. *Figure 2 : Evolution du nombre moyen de pyrale par piège selon l’année*. URL: https://www.arvalis-infos.fr/_plugins/WMS_BO_Gallery/page/getElementStream.jspz?id=72073&prop=image (visited on 10/17/2021).

[10] ARVALIS. *Jaunisse Nanisante de l’Orge (JNO) - Maladie virale sur Blé tendre, blé dur, triticale, ARVALIS Résultats 2013*. fr-FR. http://www.fiches.arvalis-infos.fr/fiche_accident/fiches_accidents.php?mode=fa&type_cul=1&type_acc=7&id_acc=53. 2013. (Visited on 10/17/2021).

[11] ARVALIS. *Protéger les semis de maïs des corbeaux et corneilles*. https://www.arvalis-infos.fr/des-solutions-a-combiner-des-le-semis-du-ma-s-@/view-21348-arvarticle.html. 2020. (Visited on 10/17/2021).

[12] ARVALIS. *Pyrale du maïs - Ravageur sur Maïs, ARVALIS Résultats 2013*. fr-FR. http://www.fiches.arvalis-infos.fr/fiche_accident/fiches_accidents.php?mode=fa&type_cul=3&type_acc=3&id_acc=126. 2013. (Visited on 10/17/2021).

[13] ARVALIS. *Taupins - Ravageur sur Blé tendre, blé dur, triticale, ARVALIS Résultats 2013*. fr-FR. http://www.fiches.arvalis-infos.fr. 2013. (Visited on 01/29/2022).

[14] Meysam Asgari-Chenaghlu et al. “Topic Detection and Tracking Techniques on Twitter: A Systematic Review”. In: *Complexity* 2021 (June 2021), pp. 1–15.

[15] Sophie Aubin et al. “Landscaping the Use of Semantics to Enhance the Interoperability of Agricultural Data”. In: 2017.

[16] Samuel Auclair et al. "SURICATE-Nat: Innovative citizen centered platform for Twitter based natural disaster monitoring". en. In: *2019 International Conference on Information and Communication Technologies for Disaster Management (ICT-DM)*. Paris, France: IEEE, Dec. 2019, pp. 1–8. ISBN: 978-1-72814-920-2. (Visited on 11/10/2021).

[17] Hussein Baalbaki et al. "KEMA: Knowledge-Graph Embedding Using Modular Arithmetic". In: *The 34th International Conference on Software Engineering and Knowledge Engineering*. 2022.

[18] Prafulla Bafna, Dhanya Pramod, and Anagha Vaidya. "Document clustering: TF-IDF approach". In: *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*. Chennai, India: IEEE, Mar. 2016, pp. 61–66. ISBN: 978-1-4673-9939-5.

[19] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate". In: *arXiv preprint arXiv:1409.0473* (2014).

[20] Iz Beltagy, Kyle Lo, and Arman Cohan. "SciBERT: A Pretrained Language Model for Scientific Text". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics, 2019, pp. 3613–3618.

[21] Tamara Ben-Ari et al. "Causes and implications of the unforeseen 2016 extreme yield loss in the breadbasket of France". en. In: *Nature Communications* 9.1 (Dec. 2018), p. 1627. (Visited on 10/17/2021).

[22] Yoshua Bengio et al. "A Neural Probabilistic Language Model". In: *J. Mach. Learn. Res.* 3 (2003), pp. 1137–1155. ISSN: 1532-4435.

[23] James Bergstra and Yoshua Bengio. "Random search for hyper-parameter optimization." In: *Journal of machine learning research* 13.2 (2012).

[24] Stéphan Bernard and Catherine Roussey. *pdf2blocks*. Version 1.1. There is a git related to this code : https://gitlab.irstea.fr/copain/pdf2blocs. Oct. 2020. DOI: 10.5281/zenodo.6605450. URL: https://doi.org/10.5281/zenodo.6605450.

[25] Tim Berners-Lee, James Hendler, and Olli Lassila. "The Semantic Web" in Scientific American". In: *Scientific American Magazine* 284 (May 2001).

[26] David M Blei, Andrew Y Ng, and Michael I Jordan. "Latent Dirichlet Allocation". In: *Journal of Machine Learning Research* 3 (2003), pp. 993–1022.

[27] Piotr Bojanowski et al. "Enriching word vectors with subword information". In: *Transactions of the association for computational linguistics* 5 (2017), pp. 135–146.

[28] Djallel Eddine Boubiche et al. "Mobile crowd sensing–Taxonomy, applications, challenges, and solutions". In: *Computers in Human Behavior* 101 (2019), pp. 352–370.

[29] Kendrick Boyd et al. "Area under the Precision-Recall Curve: Point Estimates and Confidence Intervals". In: *Advanced Information Systems Engineering*. Vol. 7908. Series Title: Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 451–466.

[30] Claudia Breazzano, Danilo Croce, and Roberto Basili. "MT-GAN-BERT: Multi-Task and Generative Adversarial Learning for Sustainable Language Processing." In: *NL4AI@ AI* IA*. 2021.

[31] Caterina Caracciolo et al. "The AGROVOC linked dataset". In: *Semantic Web Journal* (Jan. 2013). DOI: 10.3233/SW-130106.

[32] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien, eds. *Semi-supervised learning*. en. Adaptive computation and machine learning. OCLC: ocm64898359. Cambridge, Mass: MIT Press, 2006. ISBN: 978-0-262-03358-9.

[33] KR Chowdhary. "Natural language processing". In: *Fundamentals of artificial intelligence* (2020), pp. 603–649.

[34] Junyoung Chung et al. "Empirical evaluation of gated recurrent neural networks on sequence modeling". In: *arXiv preprint arXiv:1412.3555* (2014).

[35] Sidney Cox. "Information technology: the global key to precision agriculture and sustainability". In: *Computers and electronics in agriculture* 36.2-3 (2002), pp. 93–111.

[36] Danilo Croce, Giuseppe Castellucci, and Roberto Basili. "GAN-BERT: Generative Adversarial Learning for Robust Text Classification with a Bunch of Labeled Examples". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, 2020, pp. 2114–2119. DOI: 10.18653/v1/2020.acl-main.191.

[37] Benjamin Danielsson et al. "Classifying Implant-Bearing Patients via their Medical Histories: a Pre-Study on Swedish EMRs with Semi-Supervised GAN-BERT". In: *Proceedings of the 13th LREC Conference (LREC2022), Marseille, France*. 2022, pp. 21–23.

[38] Christian W. Dawson and Robert Wilby. "An artificial neural network approach to rainfall-runoff modelling". In: *Hydrological Sciences Journal* 43.1 (Feb. 1998), pp. 47–66. ISSN: 0262-6667, 2150-3435. DOI: 10.1080/02626669809492102.

[39] Matthieu De Clercq, Anshu Vats, and Alvaro Biel. "Agriculture 4.0: The future of farming technology". In: *Proceedings of the World Government Summit, Dubai, UAE* (2018), pp. 11–13.

[40] *Definition of Interoperability*. http://interoperability-definition.info/en/. Accessed: 2021-01-30.

[41] Tine Defour. *EIP-AGRI Brochure Agricultural Knowledge and Innovation Systems*. Text. Feb. 2018. URL: https://ec.europa.eu/eip/agriculture/en/publications/eip-agri-brochure-agricultural-knowledge-and (visited on 06/07/2021).

[42] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019. arXiv: 1810.04805 [cs.CL].

[43] Dejing Dou, Hao Wang, and Haishan Liu. "Semantic Data Mining: A Survey of Ontology-based Approaches". In: Feb. 2015.

[44] Brett Drury and M. Roche. "A survey of the applications of text mining for agriculture". In: *Comput. Electron. Agric.* 163 (2019).

[45] Brett Drury et al. "A survey of semantic web technology for agriculture". In: *Information Processing in Agriculture* 6.4 (2019), pp. 487–501. ISSN: 2214-3173. DOI: 10.1016/j.inpa.2019.02.001. URL: http://www.sciencedirect.com/science/article/pii/S2214317318302580.

[46] Yifan Du. "Collaborative Crowdsensing at the Edge". Theses. Sorbonne Université, July 2020.

[47] Kawin Ethayarajh. "How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2019, pp. 55–65.

[48] European and Mediterranean Plant Protection Organization. *EPPO Global Database*. (available online). 2022. URL: https://gd.eppo.int.

[49] Jérôme Euzenat and Pavel Shvaiko. *Ontology matching*. 2nd. Heidelberg (DE): Springer-Verlag, 2013.

[50] Robert M Fano. "Transmission of information: A statistical theory of communications". In: *American Journal of Physics* 29.11 (1961), pp. 793–794.

[51] M. Fiorelli et al. "CODA: Computer-aided ontology development architecture". In: *IBM J. Res. Dev.* 58 (2014).

[52] Manuel Fiorelli et al. "Computer-aided Ontology Development: an integrated environment". In: *New Challenges For NLP Frameworks (workshop held in conjuction with LREC 2010)*. European Language Resources Association (ELRA). 2010, pp. 33–40.

[53] Simon Gabay et al. "From FreEM to D'AlemBERT: a Large Corpus and a Language Model for Early Modern French". In: *arXiv preprint arXiv:2202.09452* (2022).

[54] Raghu K Ganti, Fan Ye, and Hui Lei. "Mobile crowdsensing: current state and future challenges". In: *IEEE communications Magazine* 49.11 (2011), pp. 32–39.

[55] Demin Gao et al. "A Framework for Agricultural Pest and Disease Monitoring Based on Internet-of-Things and Unmanned Aerial Vehicles". In: *Sensors* 20 (2020), p. 1487.

[56] Ruiying Geng et al. "Induction Networks for Few-Shot Text Classification". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3904–3913.

[57] Cheng Hian Goh and Stuart E. Madnick. "Representing and Reasoning about Semantic Conflicts in Heterogeneous Information Systems". AAI0597943. PhD thesis. USA, 1997.

[58] Yoav Goldberg and Omer Levy. "word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method". In: *arXiv preprint arXiv:1402.3722* (2014).

[59] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. `http://www.deeplearningbook.org`. MIT Press, 2016.

[60] Ian Goodfellow et al. "Generative adversarial nets". In: *Advances in neural information processing systems* 27 (2014).

[61] Yu Gu et al. "Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing". In: *CoRR* abs/2007.15779 (2020). arXiv: `2007.15779`.

[62] Yanzhu Guo et al. "BERTweetFR : Domain Adaptation of Pre-Trained Language Models for French Tweets". In: *Proceedings of the Seventh Workshop on Noisy User-generated Text, W-NUT*. Online: Association for Computational Linguistics, 2021, pp. 445–450.

[63] Graeme Hammer et al. "Advances in application of climate prediction in agriculture". In: *Agricultural systems* 70.2-3 (2001), pp. 515–553.

[64] Courtney Heldreth et al. "What does AI mean for smallholder farmers?: a proposal for farmer-centered AI research". In: *Interactions* 28.4 (July 2021), pp. 56–60. ISSN: 1072-5520, 1558-3449.

[65] Martinand others Heusel. "Gans trained by a two time-scale update rule converge to a local nash equilibrium". In: *Advances in neural information processing systems* 30 (2017).

[66] Julia Hirschberg and Christopher D Manning. "Advances in natural language processing". In: *Science* 349.6245 (2015), pp. 261–266.

[67] Sepp Hochreiter and Jürgen Schmidhuber. "Long short-term memory". In: *Neural computation* 9.8 (1997), pp. 1735–1780.

[68] Mohammad Hossin and Sulaiman M.N. "A Review on Evaluation Metrics for Data Classification Evaluations". In: *International Journal of Data Mining & Knowledge Management Process* 5 (Mar. 2015), pp. 01–11. DOI: 10.5121/ijdkp.2015.5201.

[69] Jeremy Howard and Sebastian Ruder. "Universal Language Model Fine-tuning for Text Classification". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2018, pp. 328–339.

[70] Rania Ibrahim et al. "Tools and approaches for topic detection from Twitter streams: survey". In: *Knowledge and Information Systems* 54.3 (Mar. 2018), pp. 511–539.

[71] Julie Ingram. "Farmer-Scientist Knowledge Exchange". In: *Encyclopedia of Food and Agricultural Ethics*. Dordrecht: Springer Netherlands, 2014, pp. 722–729. ISBN: 978-94-007-0929-4. DOI: 10.1007/978-94-007-0929-4_68.

[72] Jasmine Irani, Nitin Pise, and Madhura Phatak. "Clustering techniques and the similarity measures used in clustering: A survey". In: *International journal of computer applications* 134.7 (2016). Publisher: Foundation of Computer Science, pp. 9–14.

[73] Devlin Jacob et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186.

[74] Meynard Jean-Marc et al. "Socio-technical lock-in hinders crop diversification in France". In: *Agronomy for Sustainable Development* 38 (Oct. 2018). DOI: 10.1007/s13593-018-0535-1.

[75] Shufan Jiang et al. "Towards the Integration of Agricultural Data from Heterogeneous Sources: Perspectives for the French Agricultural Context Using Semantic Technologies". In: *Advanced Information Systems Engineering Workshops - CAiSE 2020 International Workshops*. Vol. 382. Lecture Notes in Business Information Processing. Grenoble, France: Springer, 2020, pp. 89–94.

[76] Shufan Jiang et al. "ChouBERT: Pre-training French Language Model for Crowdsensing with Tweets in Phytosanitary Context". In: *International Conference on Research Challenges in Information Science*. Springer. 2022, pp. 653–661.

[77] Shufan Jiang et al. "Fine-tuning BERT-based models for Plant Health Bulletin Classification". In: *Technology and Environment Workshop at EGC'21*. Montpellier, France, 2021. URL: https://hal.archives-ouvertes.fr/hal-03122939.

[78] Shufan Jiang et al. *French Tweets about Plant Health*. Jan. 2022. DOI: 10.5281/zenodo.5853684. URL: https://doi.org/10.5281/zenodo.5853684.

[79] Shufan Jiang et al. " Informativeness In Twitter Textual Contents For Farmer-centric Pest Monitoring". In: *Decision Making Using AI in Energy and Sustainability*. (To appear). Turkey, 2022.

[80] Shufan Jiang et al. "Informativeness In Twitter Textual Contents For Farmer-centric Plant Health Monitoring". In: *International Conference on Pattern Recognition and Artificial Intelligence*. Springer. 2022, pp. 492–503.

[81] Shufan Jiang et al. "Informativité dans les Contenus Textuels Twitter pour la Phytosurveillance Centrée sur l'observation des Agriculteurs". In: *Conférence francophone sur l'Extraction et la Gestion des Connaissances*. 2022.

[82] Shufan Jiang et al. "Named Entity Recognition For Monitoring Plant Health Threats in Tweets: A ChouBERT Approach". In: *6th International Conference on Universal Village (IEEE UV)*. (To appear). Boston, USA, 2022.

[83] Shufan Jiang et al. *Vers la Reconstruction des Connaissances Agricoles : Perspectives de Détection des Risques Naturels à partir de Sources de Données Hétérogènes*. Extraction et Gestion des Connaissances (EGC). Poster. 2021. URL: `https://hal.archives-ouvertes.fr/hal-03066102`.

[84] Daniel Jiménez et al. "From Observation to Information: Data-Driven Understanding of on Farm Yield Variation". en. In: *PLOS ONE* 11.3 (Mar. 2016), e0150015.

[85] Karen Sparck Jones. "A statistical interpretation of term specificity and its application in retrieval". In: *Journal of documentation* (1972).

[86] Armand Joulin et al. "Fasttext. zip: Compressing text classification models". In: *arXiv preprint arXiv:1612.03651* (2016).

[87] Daniel Jurafsky and James H Martin. "Speech and language processing (draft)". In: *preparation [cited 2020 June 1] Available from: https://web. stanford. edu/˜ jurafsky/slp3* (2018).

[88] Pierre Karapetiantz, Agnès Lillo-Le Louët, and Cédric Bousquet. "Informativité des forums de discussion français pour l'évaluation des effets indésirables du baclofène". In: *Therapies* 74.6 (Dec. 2019), pp. 569–578.

[89] Ursula Kenny and Aine Regan. "Co-designing a smartphone app for and with farmers: Empathising with end-users' values and needs". In: *Journal of Rural Studies* 82 (Feb. 2021), pp. 148–160.

[90] Laurens Klerkx, Emma Jakku, and Pierre Labarthe. "A review of social science on digital agriculture, smart farming and agriculture 4.0: New contributions and a future research agenda". In: *NJAS - Wageningen Journal of Life Sciences* 90-91 (2019), p. 100315.

[91] Jan-Christoph Klie et al. "The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation". In: *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, 2018, pp. 5–9. URL: http://tubiblio.ulb.tu-darmstadt.de/106270/.

[92] Trupti Kodinariya and Prashant Makwana. "Review on Determining of Cluster in K-means Clustering". In: *International Journal of Advance Research in Computer Science and Management Studies* 1 (Jan. 2013), pp. 90–95.

[93] Kamran at al. Kowsari. "Text Classification Algorithms: A Survey". In: *Information* 10.4 (Apr. 2019). arXiv: 1904.08067, p. 150.

[94] Abdelghani Laifa, Laurent Gautier, and Christophe Cruz. "Impact of Textual Data Augmentation on Linguistic Pattern Extraction to Improve the Idiomaticity of Extractive Summaries". en. In: *Big Data Analytics and Knowledge Discovery*. Vol. 12925. Series Title: Lecture Notes in Computer Science. Cham: Springer, 2021, pp. 143–151.

[95] Hang Le et al. "FlauBERT: Unsupervised Language Model Pre-training for French". In: *Proceedings of The 12th Language Resources and Evaluation Conference, LREC*. Marseille, France: European Language Resources Association, 2020, pp. 2479–2490.

[96] Bandy X Lee et al. "Transforming our world: implementing the 2030 agenda through sustainable development goal indicators". In: *Journal of public health policy* 37.1 (2016), pp. 13–31.

[97] Jinhyuk Lee et al. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining". In: *Bioinformatics* (2019).

[98] Mike Lewis et al. "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension". In: *arXiv preprint arXiv:1910.13461* (2019).

[99] Pengfei Liu et al. "Pre-Train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing". In: *ACM Comput. Surv.* (2022). ISSN: 0360-0300. DOI: 10.1145/3560815.

[100] Ilya Loshchilov and Frank Hutter. "Decoupled weight decay regularization". In: *arXiv preprint arXiv:1711.05101* (2017).

[101] SK Lowder, MV Sánchez, R Bertini, et al. "Farms, family farms, farmland distribution and farm labour: what do we know today?" In: *FAO Agricultural Development Economics Working Paper* (2019).

[102] Louis Martin et al. "CamemBERT: a Tasty French Language Model". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, 2020, pp. 7203–7219.

[103] Jose L Martinez-Rodriguez, Aidan Hogan, and Ivan Lopez-Arevalo. "Information extraction meets the semantic web: a survey". In: *Semantic Web* Preprint (2020), pp. 1–81.

[104] Jose L. Martinez-Rodriguez, Ivan Lopez-Arevalo, and Ana B. Rios-Alvarado. "OpenIE-based approach for Knowledge Graph construction from text". In: *Expert Systems with Applications* 113 (2018), pp. 339–355. ISSN: 0957-4174. DOI: 10.1016/j.eswa.2018.07.017. URL: https://www.sciencedirect.com/science/article/pii/S0957417418304329.

[105] Daniel Martini, M. Schmitz, and E. Mietzsch. "agroRDF as a Semantic Overlay to agroXML : a General Model for Enhancing Interoperability in Agrifood Data Standards". In: 2013.

[106] Luca Matteis et al. "Crop Ontology: Vocabulary For Crop-related Concepts". In: *Semantics for Biodiversity (S4BioDiv 2013)* (2013), p. 37.

[107] Warren S McCulloch and Walter Pitts. "A logical calculus of the ideas immanent in nervous activity". In: *The bulletin of mathematical biophysics* 5.4 (1943), pp. 115–133.

[108] Jorge Mendes et al. "Smartphone applications targeting precision agriculture practices—a systematic review". In: *Agronomy* 10.6 (2020), p. 855.

[109] Amil Merchant et al. "What Happens To BERT Embeddings During Fine-tuning?" In: *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*. Online: Association for Computational Linguistics, 2020, pp. 33–44.

[110] Stephen Merity, Nitish Shirish Keskar, and Richard Socher. "Regularizing and Optimizing LSTM Language Models". In: *International Conference on Learning Representations*. 2018.

[111] Lucie Michel. "Mieux valoriser les réseaux d'épidémiosurveillance lors de l'élaboration du Bulletin de Santé du Végétal". Theses. Institut des Sciences et Industries du Vivant et de l'Environnement (AgroParisTech), May 2016. URL: `https://tel.archives-ouvertes.fr/tel-01374036`.

[112] Tomas Mikolov et al. "Efficient estimation of word representations in vector space". In: *arXiv preprint arXiv:1301.3781* (2013).

[113] Jane Mills et al. "The use of Twitter for knowledge exchange on sustainable soil management". In: *Soil use and management* 35.1 (2019), pp. 195–203.

[114] Mehdi Mirza and Simon Osindero. *Conditional Generative Adversarial Nets*. Tech. rep. arXiv:1411.1784. arXiv, Nov. 2014. (Visited on 08/04/2022).

[115] Tom M. Mitchell. *Machine Learning*. McGraw Hill, 1997. ISBN: 0070428077.

[116] Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. "On the Stability of Fine-tuning {BERT}: Misconceptions, Explanations, and Strong Baselines". In: *International Conference on Learning Representations*. 2021. URL: `https://openreview.net/forum?id=nzpLWnVAyah`.

[117] Stephen Mutuvi et al. "Multilingual Epidemic Event Extraction". In: *International Conference on Asian Digital Libraries*. Springer. 2021, pp. 139–156.

[118] Joshua J. Myszewski et al. "Validating GAN-BioBERT: A Methodology for Assessing Reporting Trends in Clinical Trials". In: *Frontiers in Digital Health* 4 (May 2022), p. 878369. ISSN: 2673-253X.

[119] Rabiu Olatinwo and Gerrit Hoogenboom. "Chapter 4 - Weather-based Pest Forecasting for Efficient Crop Protection". In: *Integrated Pest Management*. Ed. by Dharam P. Abrol. San Diego: Academic Press, 2014, pp. 59–78. ISBN: 978-0-12-398529-3.

[120] Sarah Orsini. "Valoriser et expliciter les traités d'agriculture de l'Antiquité: les humanités numériques dans le projet AgroCCol". In: *Colloque Humanistica 2020*. 2020.

[121] Rachid Ouaret et al. "Random Forest Location Prediction from Social Networks during Disaster Events". en. In: *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*. Granada, Spain: IEEE, Oct. 2019, pp. 535–540.

[122] Ondřej Pánek. "Integration of Heterogeneous Data Sources Based on a Catalog of Master Entities". PhD thesis. 2015.

[123] VC Patil et al. "Internet of things (Iot) and cloud computing for agriculture: An overview". In: *Proceedings of agro-informatics and precision agriculture (AIPA 2012), India* (2012), pp. 292–296.

[124] Diego Inácio Patrício and Rafael Rieder. "Computer vision and artificial intelligence in precision agriculture for grain crops: A systematic review". In: *Computers and electronics in agriculture* 153 (2018), pp. 69–81.

[125] MT Pazienza, Armando Stellato, and Andrea Turbati. "Pearl: Projection of annotations rule language, a language for projecting (uima) annotations over rdf knowledge bases". In: (2012).

[126] Fabian Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[127] Jeffrey Pennington, Richard Socher, and Christopher D Manning. "Glove: Global vectors for word representation". In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014, pp. 1532–1543.

[128] Valeria Pesce et al. "Setting up a Global Linked Data Catalog of Datasets for Agriculture". In: *MTSR*. 2015.

[129] Matthew E. Peters et al. "Deep Contextualized Word Representations". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, 2018, pp. 2227–2237.

[130] Fabio Petroni et al. "Language Models as Knowledge Bases?" In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, 2019, pp. 2463–2473.

[131] T Phillips, L Klerkx, and M McEntee. "An investigation of social media's roles in knowledge exchange by farmers." In: *13th European International Farming Systems Association (IFSA) Symposium, Farming systems: facing uncertainties and enhancing opportunities, 1-5 July 2018, Chania, Crete, Greece*. International Farming Systems Association (IFSA) Europe. 2018, pp. 1–20.

[132] Telmo Pires, Eva Schlinger, and Dan Garrette. "How multilingual is Multilingual BERT?" In: *CoRR* abs/1906.01502 (2019). arXiv: 1906.01502. URL: http://arxiv.org/abs/1906.01502.

[133] Magali Prost, Lorène Prost, and Marianne Cerf. "Les échanges virtuels entre agriculteurs : un soutien à leurs transitions professionnelles ?" fr. In: *Raisons éducatives* 21.1 (2017), p. 129. ISSN: 1375-4459. DOI: 10.3917/raised.021.0129. URL: http://www.cairn.info/revue-raisons-educatives-2017-1-page-129.htm.

[134] Zhao Qing et al. "A pest sexual attraction monitoring system based on IoT and image processing". en. In: *Journal of Physics: Conference Series* 2005.1 (Aug. 2021), p. 012050. ISSN: 1742-6588, 1742-6596.

[135] M Atif Qureshi and Derek Greene. "Eve: explainable vector based embedding technique using wikipedia". In: *Journal of Intelligent Information Systems* 53.1 (2019), pp. 137–165.

[136] Alec Radford et al. "Improving language understanding by generative pre-training". In: *OpenAI blog* (2018).

[137] Alec Radford et al. "Language models are unsupervised multitask learners". In: *OpenAI blog* 1.8 (2019), p. 9.

[138] Anand Rajaraman and Jeffrey David Ullman. "Data Mining". In: *Mining of Massive Datasets*. Cambridge University Press, 2011, pp. 1–17. DOI: 10.1017/CBO9781139058452.002.

[139] Payam Refaeilzadeh, Lei Tang, and Huan Liu. "Cross-Validation". en. In: *Encyclopedia of Database Systems*. New York, NY: Springer New York, 2016, pp. 1–7.

[140] Adam Roberts, Colin Raffel, and Noam Shazeer. "How Much Knowledge Can You Pack Into the Parameters of a Language Model?" In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, 2020, pp. 5418–5426.

[141] Miguel Rodríguez-García and Francisco Garcia-Sanchez. "CropPestO: An Ontology Model for Identifying and Managing Plant Pests and Diseases". In: Oct. 2020, pp. 18–29. ISBN: 978-3-030-62014-1. DOI: 10.1007/978-3-030-62015-8_2.

[142] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. "A Primer in BERTology: What We Know About How BERT Works". In: *Trans. Assoc. Comput. Linguistics* 8 (2020), pp. 842–866.

[143] Francis Rousseaux. "BIG DATA and data-driven intelligent predictive algorithms to support creativity in industrial engineering". In: *Computers & Industrial Engineering* 112 (2017), pp. 459–465.

[144] Francis Rousseaux and Stéphane Cormier. "Knowledge acquisition at the time of Big Data". In: *2016 Federated Conference on Computer Science and Information Systems (FedCSIS)*. 2016, pp. 1343–1348.

[145] Catherine Roussey et al. "A methodology for the publication of agricultural alert bulletins as LOD". In: *Computers and Electronics in Agriculture* 142 (2017), pp. 632–650.

[146] Catherine Roussey. *French Crop Usage*. Version DRAFT VERSION. INRAE, 2021. URL: https://doi.org/10.15454/QHFTMX.

[147] Catherine Roussey and Stephan Bernard. "Améliorer la qualité d'un thésaurus à l'aide de requêtes SPARQL". In: *les actes du 9es atelier Recherche d'Information SEmantique (RISE 2017)*. Caen, July 2017, p. 11. URL: https://irsteadoc.irstea.fr/exl-php/document-affiche/p_recherche_publication/OUVRE_DOC/49633?fic=2017/cf2017-pub00055991.pdf.

[148] Catherine Roussey et al. "A methodology for the publication of agricultural alert bulletins as LOD". In: *Computers and Electronics in Agriculture* 142 (2017), pp. 632–650. ISSN: 0168-1699.

[149] Catherine Roussey et al. "Ontologies in Agriculture". In: *AgEng 2010, International Conference on Agricultural Engineering*. Clermont-Ferrand, France: Cemagref, Sept. 2010, p.–p. URL: https://hal.archives-ouvertes.fr/hal-00523508.

[150] Jennifer Rowley. "The wisdom hierarchy: representations of the DIKW hierarchy". In: *Journal of Information Science* 33.2 (2007), pp. 163–180. DOI: 10.1177/0165551506070706.

[151] Harald Sack and Mehwish Alam. "5.3 RDF and OWL Knowledge Graphs". In: *Knowledge Graphs*. 2020. URL: https://open.hpi.de/courses/knowledgegraphs2020/items/6GLVXmUiQ79Pf1ikqhPmOz.

[152] Takaya Saito and Marc Rehmsmeier. "The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets". In: *PLOS ONE* 10.3 (Mar. 2015), e0118432. (Visited on 11/11/2021).

[153] Tim Salimans et al. "Improved techniques for training gans". In: *Advances in neural information processing systems* 29 (2016).

[154] Victor Sanh et al. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter". In: *arXiv preprint arXiv:1910.01108* (2019).

[155] Raquel Bento Santos et al. "Semi-Supervised Annotation of Portuguese Hate Speech Across Social Media Domains". In: *11th Symposium on Languages, Applications and Technologies (SLATE 2022)*. Vol. 104. Open Access Series in Informatics (OASIcs). Dagstuhl, Germany, 2022, 11:1–11:14. ISBN: 978-3-95977-245-7.

[156] Peter Selby et al. "BrAPI—an application programming interface for plant breeding applications". In: *Bioinformatics* 35.20 (Mar. 2019), pp. 4147–4155. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btz190. eprint: https://academic.oup.com/bioinformatics/article-pdf/35/20/4147/30148301/btz190.pdf. URL: https://doi.org/10.1093/bioinformatics/btz190.

[157] Priyamvada Shankar, Christian Bitter, and Marcus Liwicki. "Digital Crop Health Monitoring by Analyzing Social Media Streams". In: *2020 IEEE / ITU International Conference on Artificial Intelligence for Good (AI4G)*. Geneva, Switzerland: IEEE, 2020, pp. 87–94.

[158] Amit P Sheth. "Changing focus on interoperability in information systems: from system, syntax, structure to semantics". In: *Interoperating geographic information systems*. Springer, 1999, pp. 5–29.

[159] Max Silberztein. *Formalizing natural languages: The NooJ approach*. John Wiley & Sons, 2016.

[160] Sonit Singh. "Natural language processing for information extraction". In: *arXiv preprint arXiv:1807.02383* (2018).

[161] Vivek Kumar Singh, Nisha Tiwari, and Shekhar Garg. "Document Clustering Using K-Means, Heuristic K-Means and Fuzzy C-Means". In: *2011 Int. Conference on Computational Intelligence and Communication Networks*. Gwalior, India: IEEE, Oct. 2011, pp. 297–301. (Visited on 07/15/2021).

[162] Michael Steinbach, George Karypis, and Vipin Kumar. *A Comparison of Document Clustering Techniques*. Report. May 2000.

[163] Douka Stella et al. "JuriBERT: A Masked-Language Model Adaptation for French Legal Text". In: *Proceedings of the Natural Legal Language Processing Workshop*. Association for Computational Linguistics, Nov. 2021, pp. 95–101.

[164] Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. "Yago: A large ontology from wikipedia and wordnet". In: *Journal of Web Semantics* 6.3 (2008), pp. 203–217.

[165] Hoang Thang Ta et al. "GAN-BERT, an Adversarial Learning Architecture for Paraphrase Identification". In: *CEUR Workshop Proceedings*. Vol. 3202. CEUR-WS. 2022.

[166] Raihan Tanvir et al. *A GAN-BERT Based Approach for Bengali Text Classification with a Few Labeled Examples*. Tech. rep. EasyChair, 2022.

[167] Bertille Thareau and Karine Daniel. "Le numérique accompagne les mutations économiques et sociales de l'agriculture". fr. In: *Sciences Eaux & Territoires* Numéro 29.3 (2019), p. 44.

[168] Elodie Thiéblin et al. "Survey on complex ontology matching". In: *Semantic Web* Preprint (2019), pp. 1–39.

[169] Brice Thollet. "MEDIAS SOCIAUX EN AGRICULTURE : Contribution à l'analyse des usages et de leur potentiel d'apprentissage pour la transition agroécologique". MA thesis. Dijon, France: AgroSup Dijon, Aug. 2020.

[170] Ba-Huy Tran et al. "A Semantic Mediator for Handling Heterogeneity of Spatio-Temporal Environment Data". In: *9th International Conference on Metadata and Semantics Research*. Vol. Volume 544. Metadata and Semantics Research Communications in Computer and Information Science. Manchester, United Kingdom, Sept. 2015, pp. 381–392. DOI: 10.1007/978-3-319-24129-6_33.

[171] Paolo Tripicchio et al. "Towards smart farming and sustainable agriculture with drones". In: *2015 International Conference on Intelligent Environments*. IEEE. 2015, pp. 140–143.

[172] K. Trivedi. *Fast-BERT*. https://github.com/kaushaltrivedi/fast-bert. Version 1.9.1. 2020.

[173] Nicolas Turenne. *reportsOCR.zip*. https://www.data.gouv.fr/fr/datasets/r/c745b0bf-b135-4dc0-ba04-1e15c1b77899.

[174] Nicolas Turenne and Mathieu Andro. *Maladies des Cultures*. INRA, LISIS, Feb. 2017. DOI: 10.5281/zenodo.268301. URL: https://doi.org/10.5281/zenodo.268301.

[175] Nicolas Turenne et al. "Open Data Platform for Knowledge Access in Plant Health Domain : VESPA Mining". In: *CoRR* abs/1504.06077 (2015). arXiv: 1504.06077.

[176] Sarah Valentin et al. "PADI-web 3.0: A new framework for extracting and disseminating fine-grained information from the news for animal disease surveillance". In: *One Health* 13 (2021), p. 100357. ISSN: 2352-7714. DOI: https://doi.org/10.1016/j.onehlt.2021.100357. URL: https://www.sciencedirect.com/science/article/pii/S2352771421001476.

[177] Sarah Valentin et al. "PADI-web: A multilingual event-based surveillance system for monitoring animal infectious diseases". In: *Computers and Electronics in Agriculture* 169 (2020), p. 105163.

[178] Dialekti Valsamou. "Extraction d'Information pour les réseaux de régulation de la graine chez Arabidopsis Thaliana." PhD thesis. Université Paris-Saclay (ComUE), 2017.

[179] Laurens Van der Maaten and Geoffrey Hinton. "Visualizing data using t-SNE." In: *Journal of machine learning research* 9.11 (2008).

[180] Jesper E Van Engelen and Holger H Hoos. "A survey on semi-supervised learning". In: *Machine Learning* 109.2 (2020), pp. 373–440.

[181] Maarten Van Steen and A Tanenbaum. "Distributed systems principles and paradigms". In: *Network* 2 (2002), p. 28.

[182] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[183] Jesse Vig. "A Multiscale Visualization of Attention in the Transformer Model". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 37–42. DOI: 10.18653/v1/P19-3007.

[184] Hanna M Wallach. "Conditional random fields: An introduction". In: *Technical Reports (CIS)* (2004), p. 22.

[185] Achim Walter et al. "Opinion: Smart farming is key to developing sustainable agriculture". In: *Proceedings of the National Academy of Sciences* 114.24 (2017), pp. 6148–6150. ISSN: 0027-8424. DOI: 10.1073/pnas.1707462114. eprint: https://www.pnas.org/content/114/24/6148.full.pdf.

[186] Dong Wang, Tarek Abdelzaher, and Lance Kaplan. *Social sensing: building reliable systems on unreliable data*. Morgan Kaufmann, 2015.

[187] Alex Warstadt et al. "Investigating BERT's Knowledge of Language: Five Analysis Methods with NPIs". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 2877–2887. DOI: 10.18653/v1/D19-1286.

[188] Andrew Weiss. "Google Ngram Viewer". In: *The Complete Guide to Using Google in Libraries: Instruction, Administration, and Staff Productivity* 1 (2015), p. 183.

[189] Marijke Welvaert et al. "Limits of use of social media for monitoring biosecurity events". In: *PloS one* 12.2 (2017).

[190] Tai Wen et al. "exBERT: Extending Pre-trained Models with Domain-specific Vocabulary Under Constrained Training Resources". In: *Findings of the Association for Computational Linguistics: EMNLP*. Online: Association for Computational Linguistics, 2020, pp. 1433–1439.

[191] Emma White et al. "Report from the conference, 'identifying obstacles to applying big data in agriculture'". In: *Precision Agriculture* 22 (July 2020). DOI: 10.1007/s11119-020-09738-y.

[192] Daya Wimalasuriya and Dejing Dou. "Ontology-based information extraction: An Introduction and a survey of current approaches". In: *J. Information Science* 36 (2010), pp. 306–323.

[193] Thomas Wolf et al. "Transformers: State-of-the-Art Natural Language Processing". In: *Proc. of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45.

[194] Sjaak Wolfert et al. "Big data in smart farming–a review". In: *Agricultural systems* 153 (2017), pp. 69–80.

[195] Yonghui Wu et al. "Google's neural machine translation system: Bridging the gap between human and machine translation". In: *arXiv preprint arXiv:1609.08144* (2016).

[196] Gang Yu et al. "Adversarial active learning for the identification of medical concepts and annotation inconsistency". In: *Journal of Biomedical Informatics* 108 (Aug. 2020), p. 103481. ISSN: 15320464. DOI: 10.1016/j.jbi.2020.103481.

[197] Hamada M. Zahera. "Fine-tuned BERT Model for Multi-Label Tweets Classification". In: *TREC*. 2019.

[198] Marcia Lei Zeng. "Interoperability". In: 46.2 (2019), pp. 122–146. ISSN: 0943-7444. DOI: 10.5771/0943-7444-2019-2-122.

[199] Li Zhang, Jun Li, and Chao Wang. "Automatic synonym extraction using Word2Vec and spectral clustering". In: *2017 36th Chinese Control Conference (CCC)*. 2017, pp. 5629–5632. DOI: 10.23919/ChiCC.2017.8028251.

[200] Yijia Zhang et al. "BioWordVec, improving biomedical word embeddings with subword information and MeSH". In: *Scientific data* 6.1 (2019), pp. 1–9.

[201] Hend Zouari. "French AXA Insurance Word Embeddings : Effects of Fine-tuning BERT and Camembert on AXA France's data". MA thesis. KTH, School of Electrical Engineering and Computer Science (EECS), 2020, p. 178.

Intégration de données textuelles pour la détection de risques naturels en agriculture

Résumé : L'agriculture entre dans l'ère numérique grâce aux données (qui ouvrent à l'agriculture de précision) ou aux connaissances (qui ouvrent à de nouveaux outils d'aide à la décision). Les technologies modernes et les dispositifs IoT ont été appliqués pour améliorer les processus agricoles. Un scénario d'application consiste à la phytosurveillance à l'aide de capteurs et de techniques d'analyse des données. Cependant, la plupart des solutions existantes basées sur des dispositifs spécifiques et des technologies d'imagerie nécessitent un investissement financier, inaccessible aux petits exploitants. L'absence de contribution des agriculteurs à la collecte des données et la prise de décision dans ces solutions soulève des problèmes de confiance entre les agriculteurs et les technologies d'agriculture intelligente. D'autre part, les données textuelles en agriculture, e.g. les échanges parmi les agriculteurs sur réseaux sociaux, peuvent être une source de connaissances. Ces connaissances ont une grande valeur lorsqu'elles sont formalisées, contextualisées et intégrées avec d'autres données. Poussée par la connectivité croissante des agriculteurs et l'émergence de communautés agricoles en ligne, cette thèse propose : (1) d'utiliser Twitter comme une plateforme ouverte de crowdsensing pour acquérir les perceptions des gens sur la santé des cultures afin que nous puissions inclure la participation des agriculteurs dans la reconstruction des connaissances agricoles. (2) d'utiliser des modèles de langage pré-entraînés comme une base de connaissances implicite et spécifique au domaine qui intègre des textes hétérogènes et soutient l'extraction d'informations du texte.

Mots clés : TALN, Intelligence Artificielle, Apprentissage Machine, Santé Végétale, Média Sociaux

Integrating textual data towards crowdsensing natural hazards in agriculture

Abstract: Agriculture is entering the digital age through data (which opens up precision agriculture) or knowledge (which opens up new decision support tools). Modern technologies and IoT devices have been applied to improve agricultural processes. One application scenario is plant monitoring using sensors and data analysis techniques. However, most existing solutions based on specific devices and imaging technologies require a financial investment, which is inaccessible to small farmers. Furthermore, the lack of farmer input into data collection and decision-making in these solutions raises trust issues between farmers and smart farming technologies. On the other hand, textual data in agriculture, e.g. exchanges among farmers on social networks, can be a source of knowledge. This knowledge has great value when it is formalized, contextualized and integrated with other data. Crowdsensing is a sensing paradigm that allows ordinary people to contribute with data that their mobile devices equipped with sensors collect or generate. Farmers' observations reflect their knowledge and experience in plant health monitoring. Driven by the increasing connectivity of farmers and the emergence of online farming communities, this thesis proposes: (1) to use Twitter as an open crowdsensing platform to acquire people's perceptions of crop health so that we can include farmer participation in agricultural knowledge reconstruction. (2) to use pre-trained language models as an implicit and domain-specific knowledge base that integrates heterogeneous texts and supports information extraction from text.

Keywords : NLP, Artificial Intelligence, Machine Learning, Plant Health, Social Media

---

Discipline : Informatique

Spécialité : Agriculture numérique de précision, Traitement Automatique du Langage Naturel, Apprentissage Machine et Intelligence Artificielle

Université de Reims Champagne-Ardenne
Centre de Recherche en STIC (CReSTIC, EA 3804),
UFR Sciences Exactes et Naturelles
Moulin de la Housse, BP 1039, 51687 Reims, France