

Fine-tuning BERT-based models for Plant Health Bulletin Classification



Shufan Jiang^{1,2}, Rafael Angarita¹, Stéphane Cormier², Francis Rousseaux²

1. Institut Supérieur d'Electronique de Paris, LISITE, Paris, France name.lastname@isep.fr,

2. Université de Reims Champagne Ardenne, CReSTIC, Reims, France name.lastname@univ-reims.fr

Introduction - Context of our research



Unstructured data from Twitter

LA GESTION DU VERGER CIDRICOLE

J'ai repris 1,3 ha en production et replanté 4 ha. Sur ces nouvelles plantations, l'arrosage est la seule intervention. Dans le verger en production, je passe le broyeur entre les arbres, je taille les branches basses, j'enlève le gui et les ...

Unstructured data from farmers experiences

★ **Charançon de la tige du colza**

Cette semaine, les piégeages significatifs (> à 5 individus/cuvette) ne sont signalés que pour deux parcelles (dans le Gers). **On retrouve en moyenne 2 charançons de la tige du colza dans les cuvettes (contre 4 individus en moyenne la semaine dernière). Les captures sont de moins en moins importantes et l'on se dirige vers la fin du vol.**

Le vol du charançon de la tige du colza a démarré de façon intense et regroupé il y a maintenant trois semaines. Les conditions météorologiques lui ont été très favorables depuis.

Attention toutefois, on retrouve également du charançon de la tige du chou, non nuisible pour le colza dans tous les départements où l'on voit du charançon de la tige du colza (voir encadré ci-dessous pour éviter la confusion entre les deux charançons). Attention toutefois, on retrouve également du charançon de la tige du chou, non nuisible pour le colza dans tous les départements où l'on voit du charançon de la tige du colza (voir encadré ci-dessous pour éviter la confusion entre les deux charançons).



Dégât engendré par le charançon de la tige du colza (Photo Terres Inovia)

BULLETIN de SANTÉ du VÉGÉTAL
ÉCOPHYTO

Bulletin de Santé du Végétal Nouvelle-Aquitaine / Edition Aquitaine
Grandes cultures - N°04 du 27 février 2020

2 / 10

Semi-structured data the French Plants Health Bulletins

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

Structured data from a weather sensor

Introduction - Plant Health Bulletins (BSV)



BULLETIN DE SANTE DU VEGETAL

MIDI-PYRENEES

Grandes Cultures - n°13

25 mars 2010

A retenir

CEREALES A PAILLE

Pietin verse : risque parfois élevé pour les semis d'octobre dans les limons et les rotations avec retour fréquent de blé.

Septoriose : risque élevé pour les semis d'octobre, particulièrement pour les variétés de blé tendre sensibles, semées mi-octobre.

Helminthosporiose : surveillez les semis précoces de variétés d'orge sensibles.

PROTEAGINEUX

Sitones : attaques favorisées par le redoux des températures, seuil de nuisibilité atteint dans certaines situations. Soyez vigilant, en particulier pour les pois de printemps semés en janvier – février.

COLZA

Charançon de la tige du colza : toujours en activité sur l'ensemble de la région. Période de risque et seuil de nuisibilité atteints sur les parcelles du réseau non protégées au cours des 15 derniers jours.

Meligèthe : aucun risque à ce jour. Surveillez très régulièrement vos parcelles.

CÉRÉALES À PAILLE

• Stades phénologiques
Semis du 15/10 au 30/10 : Sur notre réseau, les blés durs sont entre le stade épi 1 cm et 2 nœuds. Les blés tendres sont entre le début de la montaison et épi 1cm.
Semis du 30/10 au 20/11 : les blés durs sont entre le début montaison et épi 1cm voir 1 nœud pour les variétés les plus précoces. Les blés tendres sont entre fin tallage et épi 1cm.
Les blés semés au 15/12 sont à plein tallage.
Les orges sont entre le stade fin tallage et début montaison pour les semis d'octobre et de novembre. Les semis de décembre sont au stade plein tallage.

• Dégâts de froid
Le dernier épisode de froid a pu entraîner des gels d'épis sur les blés durs les plus avancés (au stade 1-2 nœuds) pouvant toucher le maître brin et les talles. C'est dans ces situations que les gels d'épis pourront avoir une incidence sur les densités épis et le potentiel de rendement.

Directeur de publication : Jean-Louis CAZAUBON
Président de la Chambre Régionale d'Agriculture de Midi-Pyrénées
BP 22107 - 31221 CASTANET TOULOUSAIN Cx
Tél 05 61 75 20 50 - Fax 05 61 75 16 60

Dépôt légal : à parution
ISSN en cours

BULLETIN DE SANTÉ DU VÉGÉTAL – GRANDES CULTURES N° 13 DU 25 MARS 2010 – Page 1/5



Bulletin de santé du végétal

Oléagineux

du 06/04/2011 au 12/04/2011

N° 22

COLZA

RESEAU 2010 - 2011

Les observations ont été réalisées sur 103 parcelles au cours des derniers jours y compris les parcelles mise en place spécifiquement pour la réalisation des kits pétioles sclerotinia.

STADE DES COLZAS

Plus de 60 % des parcelles du réseau ont atteint le stade G1 voir l'ont dépassé.

SCLEROTINIA

Contexte d'observations

La quasi-totalité des parcelles ont atteint la période de risque sclerotinia. Les parcelles les plus précoces ont déjà atteint cette période la semaine précédente.
La durée de floraison définira la longueur de la période de risque.

Kits Pétioles :
41 kits pétioles ont été réalisés depuis le 28/03/2011. Les résultats de 40 kits sont disponibles (cf. carte en annexe). Les taux de contamination sont compris entre 12 et 100 %. La moyenne est de 67 %.

Période de risque
Le stade G1 est le stade de début de la période de risque, le stade G1 correspond aux 10 premières silques formées sur les hampes principales (longueur inférieure à 2 cm).

A la chute des pétioles sur les feuilles (stade G1) et en conditions optimales, le champignon pourra coloniser la feuille puis la tige du colza. Attention, la date de ce stade peut varier d'une parcelle à l'autre.

Seuil de nuisibilité
Il n'existe pas pour le sclerotinia du colza de seuil de nuisibilité étant donné que la protection est préventive.
Cependant le niveau de risque peut être évalué selon :
- le pourcentage de pétioles contaminées (X4 pétioles) : risque avéré au-delà de 30 %.
- le nombre de cultures sensibles dans la rotation,
- les attaques les années antérieures sur la parcelle,
- les conditions climatiques humides au mois de mars favorables à la germination des sclérotes.

Ensuite, le climat durant toute la floraison favorisera ou non l'expression de la maladie : humidité relative de plus de 90 % dans le couvert durant 3 jours pendant la floraison et une température moyenne journalière supérieure à 10°C.

CHARANÇON DES SILQUES

Contexte d'observations
Près de la moitié des 72 parcelles observées signalait la présence de charançon des silques. Les températures chaudes des derniers jours ont été très favorables aux vols (températures supérieures à 17 °C).
Pour l'instant plus de 85 % des parcelles n'ont pas atteint le seuil de nuisibilité.

Directeur de publication : Jean-Pierre LEBAILLARD, Président de la Chambre régionale d'Agriculture de Centre
12 avenue des Brûlés de l'Homme - 41000 ORLÉANS

Ce bulletin est produit à partir d'observations personnelles. Il donne une tendance de la situation sanitaire régionale, qui ne peut pas être transposée telle quelle à la parcelle.
La Chambre Régionale d'Agriculture de Centre engage ses responsables quant aux décisions prises par les agriculteurs pour la protection de leurs cultures.

ecophyto2018
un partenariat pour mieux gérer les risques

Seuil de nuisibilité : à partir de 2 nœuds des blés si plus de 20% des troisièmes feuilles présentent des symptômes.

• Rouille brune : Aucun symptôme n'a été observé sur notre réseau. A ce jour, le risque rouille brune est faible quelle que soit la date de semis. L'arrivée de la rouille brune devrait être assez tardive. Ceci s'explique par un niveau d'innoculum bas en fin d'été (été sec et peu de repousses de blé) et par un hiver froid qui n'a pas permis de multiplier les contaminations primaires d'automne.

• Seuil de nuisibilité : apparition des pustules sur l'une des 3 feuilles supérieures.

• Helminthosporiose sur orge : Des symptômes importants d'Helminthosporiose ont été observés sur les feuilles basses sur les variétés sensibles semées précocement (notamment Kétos et Azurel).

Évaluation du risque : Surveillez les semis précoces de variétés d'orge sensibles.

• Seuil de nuisibilité : apparition des premiers symptômes sur l'une des 3 feuilles supérieures.

• Rouille naine sur orge : quelques pustules de rouille naine ont été observées sur des semis d'octobre d'orge dans le Gers.

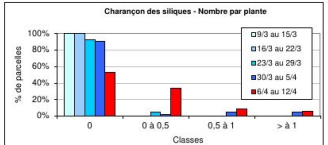
Évaluation du risque : Le développement de cette maladie peut être explosif. Surveillez régulièrement vos parcelles.

Directeur de publication : Jean-Pierre LEBAILLARD, Président de la Chambre régionale d'Agriculture de Centre
12 avenue des Brûlés de l'Homme - 41000 ORLÉANS

Ce bulletin est produit à partir d'observations personnelles. Il donne une tendance de la situation sanitaire régionale, qui ne peut pas être transposée telle quelle à la parcelle.
La Chambre Régionale d'Agriculture de Centre engage ses responsables quant aux décisions prises par les agriculteurs pour la protection de leurs cultures.

ecophyto2018
un partenariat pour mieux gérer les risques

Charançon des silques - Nombre par plante




Directeur de publication : Jean-Pierre LEBAILLARD, Président de la Chambre régionale d'Agriculture de Centre
12 avenue des Brûlés de l'Homme - 41000 ORLÉANS

Ce bulletin est produit à partir d'observations personnelles. Il donne une tendance de la situation sanitaire régionale, qui ne peut pas être transposée telle quelle à la parcelle.
La Chambre Régionale d'Agriculture de Centre engage ses responsables quant aux décisions prises par les agriculteurs pour la protection de leurs cultures.

ecophyto2018
un partenariat pour mieux gérer les risques


BULLETIN DE SANTÉ DU VÉGÉTAL – GRANDES CULTURES N° 13 DU 25 MARS 2010 – Page 2/5



Introduction - Text Classification & PestObserver Site

**PESTOBSERVER**

Mon profil Déconnexion

 Plante

 Maladie

 Ravageur
pyrale du jasmin

Date de début
00/00/0000

Date de fin
07/12/2017

→ LANCER LA RECHERCHE

Les Bulletins

Grandes Cultures

Années

Toutes

1900 (1)

2013 (10)

2014 (7)

2016 (3)

21/21

Trier : Nom Région Date

PROVENCE-ALPES-CÔT... 01/01/1900

draaf.paca.agriculture.gouv.fr_im...

CORSE 24/07/2013

draaf.corse.agriculture.gouv.fr_i...

CORSE 01/08/2013

draaf.corse.agriculture.gouv.fr_i...

CORSE 13/08/2013

draaf.corse.agriculture.gouv.fr_i...

CORSE 21/08/2013

draaf.corse.agriculture.gouv.fr_i...

CORSE 11/09/2013

draaf.corse.agriculture.gouv.fr_i...

AQUITAINE 02-10-2013

bsv_n_11-pä@piniä~re-2013.10.02

AQUITAINE 02/10/2013

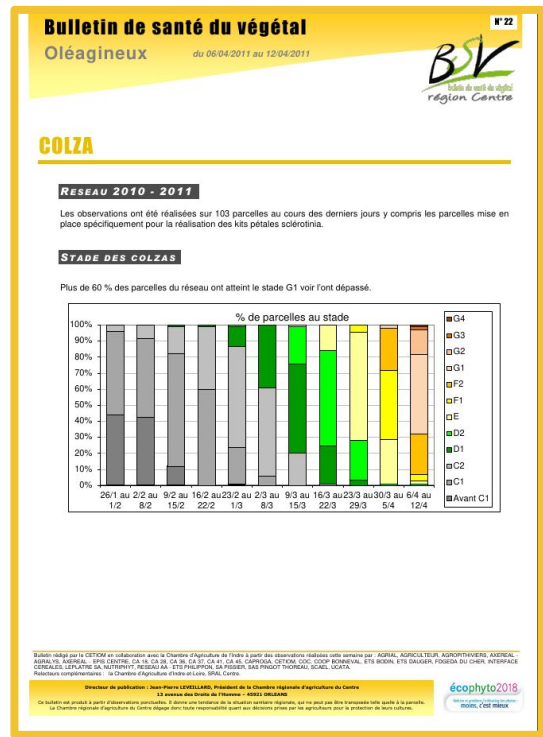
bsv_n011-neniniere-2013.10.02

Bulletins citant  pyrale du jasmin du 00/00/0000 au 07/12/2017



Existing Dataset A) Optical Character Recognition(OCR) Processed BSV

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



Bulletin de santé du végétal

N° 22
BSV Céréales à paille N° 00

Oléagineux

du 06/04/2011 au 12/04/2011

1

COLZA

RESEAU 2010 - 2011

Les observations ont été réalisées sur 103 parcelles au cours des derniers jours y compris les parcelles mise en place spécifiquement pour la réalisation des kits pétales sclérötinia.

STADE DES COLZAS

Plus de 60 % des parcelles du réseau ont atteint le stade G1 voir l'ont dépassé.

% de parcelles au stade

G4

G3

G2

G1

F2

F1

E

D2

D1

C2

C1	
26/1 au	2
2/2 au	
1/2	
8/2	
9/2 au	
15/2	
16/2 au	
23/2 au	
2/3 au	3
22/2	
1/3	
8/3	
9/3 au	
15/3	100%
16/3 au	90%
23/3 au	80%
30/3 au	70%
22/3	60%
29/3	50%
5/4	40%
6/4 au	30%
12/4	20%
Avant C1	10%
	0%
Bulletin rédigé par le CETIOM en collaboration avec la Chambre d'Agriculture de l'Indre à partir des observations réalisées cette semaine par : AGRIAL, AGRICULTEUR, AGROPITHIVIERS, AXEREA, AGRALYS, AXEREA - EPIS CENTRE, CA 18, CA 28, CA 36, CA 37, CA 41, CA 45, CAPROGA, CETIOM, COC, COOP BONNEVAL, ETS BODIN, ETS DAUGER, FDGEDA DU CHER, INTERFACE CEREALES, LEPLATRE SA, NUTRIPHYT, RESEAU AA - ETS PHILIPPON, SA PISSIER, SAS PINGOT THOREAU, SCAEL, UCATA. Relecteurs complémentaires : la Chambre d'Agriculture d'Indre-et-Loire, SRAL Centre.	

Existing Dataset A) Former BSV in XML(TEI)

SPV MINISTÈRE DE L'AGRICULTURE
DLP 21-5-85515404 (R)
Avertissements agricoles BRETAGNE
SERVICE DE LA PROTECTION DES VÉGÉTAUX
230, rue de Fougères, 35000 RENNES ☎ (05) 36 01 74
BULLETIN TECHNIQUE DE LA STATION D'AVERTISSEMENTS AGRICOLES
Publication périodique
EDITION : CULTURES MARAÎCHÈRES, LÉGUMIÈRES ET POMMES DE TERRE
BULLETIN N° 110 - 30 avril 1985

- Mouche de la carotte (suite)
- Serres : Attention aux ravageurs
- Pommes de terre de primeur : pas de mildiou actuellement, mais...

MOUCHE OU "VER" DE LA CAROTTE
(suite du texte paru dans le bulletin n° 109)

B/ - Traitements en végétation
Certains types de cultures de carottes font couramment appel à ce type de traitement. L'efficacité semble variable suivant les périodes d'application; les résultats sont souvent aléatoires en traitement d'automne.

Pour obtenir une efficacité maximale, tout en diminuant les risques de présence de résidus, il importe de bien suivre les conseils suivants :

- Utiliser un fort volume de bouillie à l'ha : 1 000 à 1 500 litres, afin que le produit descende jusqu'au sol.
- Pour éviter l'effet d'accumulation des produits dans la racine, ne pas employer le même produit qu'au semis.
- La dose d'emploi ne doit pas être confondue avec celle qui est utilisée au semis.

Produits homologués contre la mouche, en cours de végétation

- Chlorfenvinphos (60 g/ha) soit 1,5 l/ha de Birlane CE 40
- diéthion (75 g/ha) soit 1,5 l/ha de Rhodocide

REMARQUE: Les produits homologués contre les pucerons de la carotte peuvent aussi être employés contre la mouche, sous la responsabilité de l'utilisateur, et sans aucune garantie d'efficacité.

Délais d'utilisation avant récolte

Bromophos, chlorfenvinphos et deltaméthrine peuvent être employés jusqu'à un mois de la récolte.

Pour les autres produits, du fait du manque de références expérimentales, il convient d'être prudent et de considérer ce délai d'un mois comme étant un minimum.

IV - PROGRAMME DE TRAITEMENTS PAR TYPE DE CULTURES

A/ - Carotte d'été

- Période de récolte : juillet-août

Ce type de culture n'étant exposé qu'aux dégâts dus au premier vol, le traitement du sol suffit à protéger la culture.

B/ - Carotte d'automne-hiver

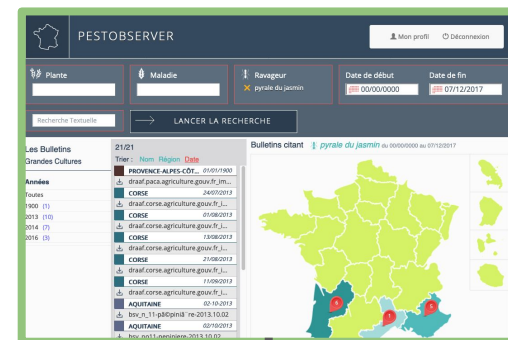
- Semis : mai-juin
- Récolte : septembre à mai de l'année suivante

ABONNEMENT 1^{er} Janv. : 31 325 F
C.C.P. 9404-94 Y Rennes - S / Rég. de Recettes de la DDA - Prix Vagabond
Service de la Protection des Végétaux - Rennes 1985
Toute reproduction même partielle est soumise à notre autorisation

```
<?xml version="1.0" encoding="UTF-8"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title>Avertissements agricoles BULLETIN N° 110 - 30 avril 1985</title>
      </titleStmt>
      <publicationStmt>
        <p>Edition grandes cultures.</p>
      </publicationStmt>
      <sourceDesc>
        <p>Service Régional de la Protection des Végétaux</p>
      </sourceDesc>
    </fileDesc>
  </teiHeader>
  <text>
    <body>
      <head>BULLETIN N° 110 - 30 avril 1985</head>
      <p>- Mouche de la carotte (suite)</p>
      <p>- Serres : Attention aux ravageurs</p>
      <p>- Pommes de terre de primeur : pas de mildiou actuellement, mais...</p>
      <div1 n="1" type="part">
        <head>MOUCHE OU "VER" DE LA CAROTTE</head>
        <p>(suite du texte paru dans le bulletin n° 109)</p>
        <div2 n="1.1" type="section">
          <head>B/ - Traitements en végétation</head>
          <p>Certains types de cultures de carottes font couramment appel à ce type de traitement. L'efficacité semble variable suivant les périodes d'application; les résultats sont souvent aléatoires en traitement d'automne.</p>
          <p>Pour obtenir une efficacité maximale, tout en diminuant les risques de présence de résidus, il importe de bien suivre les conseils suivants </p>
          <p>• Utiliser un fort volume de bouillie à l'ha : 1 000 à 1 500 litres, afin que le produit descende jusqu'au sol.</p>
          <p>• Pour éviter l'effet d'accumulation des produits dans la racine, ne pas employer le même produit qu'au semis.</p>
          <p>• La dose d'emploi ne doit pas être confondue avec celle qui est utilisée au semis.</p>
          <p>Produits homologués contre la mouche, en cours de végétation</p>
          <p>• Chlorfenvinphos (60 g/ha) soit 1,5 l/ha de Birlane CE 40</p>
          <p>• diéthion (75 g/ha) soit 1,5 l/ha de Rhodocide</p>
          <div3 n="1.1.1" type="section">
            <head>REMARQUES</head>
            <p>Les produits homologués contre les pucerons de la carotte peuvent aussi être employés contre la mouche, sous la responsabilité de l'utilisateur, et sans aucune garantie d'efficacité</p>
            <p>Délais d'utilisation avant récolte</p>
            <p>Bromophos, chlorfenvinphos et deltaméthrine peuvent être employés jusqu'à un mois de la récolte.</p>
            <p>Pour les autres produits, du fait du manque de références expérimentales, il convient d'être prudent et de considérer ce délai d'un mois comme étant un minimum.</p>
            <p>IV - PROGRAMME DE TRAITEMENTS PAR TYPE DE CULTURES</p>
            <p>A/ - Carotte d'été o Période de récolte : juillet-août</p>
            <p>Ce type de culture n'étant exposé qu'aux dégâts dus au premier vol, le traitement du sol suffit à protéger la culture.</p>
            <p>B/ - Carotte d'automne-hiver</p>
            <p>Semis : mai-juin</p>
            <p>Récolte : septembre à mai de l'année suivante</p>
          </div3>
        </div2>
      </div1>
    </body>
  </text>
</TEI>
```

Existing Dataset B) PestObserver Tags

We also obtained tags for each BSV from the PestObserver site. There are 389 bioagressor and 279 disease tags and those 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 crops names.



Name	type_id	List of BSV names
nématode des crucifères	1	
puceron cendré du pommier	2	Bioagress... [481 elements]
puceron lanigère des racines de laitue	3	Bioagress... [30 elements]
balanin	4	Bioagress... [234 elements]
nématode du pois	5	Bioagress... [0 elements]
sangler	6	Bioagress... [82 elements]
puceron brun du prunier	7	Bioagress... [0 elements]
scolyte	8	Bioagress... [200 elements]
anthonome du pommier	9	Bioagress... [314 elements]
petite mineuse des feuilles d'olivier	10	Bioagress... [0 elements]
mouche des semis	11	Bioagress... [871 elements]
zygène de l'amandier	12	Bioagress... [1 element]
chenille	13	Bioagress... [5590 elements]
thrips sur oignons	14	Bioagress... [24 elements]

Existing Dataset C) Tweets

We use concepts in FrenchCropUsage thesaurus FAIR sharing Team (2018) and the tags in item B) as filters to collect tweets for testing the classification model

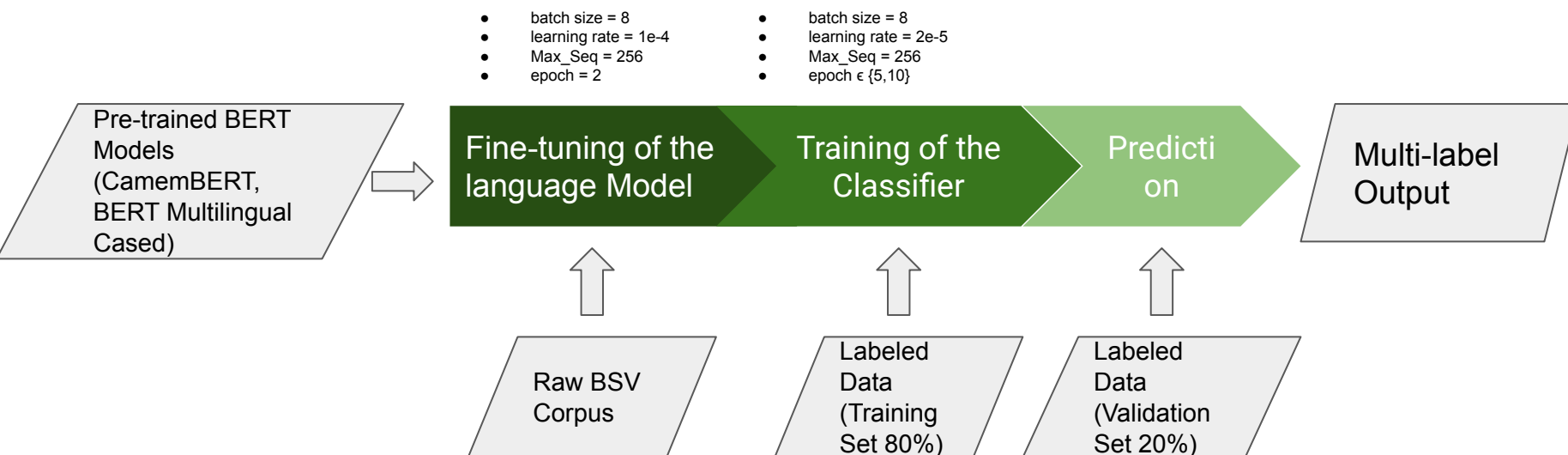
Linguistic Preprocessing for text of each BSV

We removed the following from the text of each BSV:

- Punctuation marks, URLs, phone numbers and stop words from the BSV text.
- Extra white-spaces, repeated full stops, question marks and exclamation marks.
- Continuous lines that contain less than 3 words are rows from broken tables in the original PDF file.
- Strings like "B U L L E T I N" appearing in vertical lines.

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 Wolf et al. (2020) and fast-bert Trivedi (2020).



Dataset Construction

1. For the unsupervised fine-tuning task, we extracted paragraphs from xml format BSV in item A) to make the corpus for the self-supervised fine-tuning of the language model.
2. For the classification of the topic, we randomly split 200 cleaned BSVs into 4000 chunks containing between 5 and 256 words. We classify each chunk as bioagressors and diseases according to the tags of its corresponding BSV -see item C) The length distribution and the label counts by hazard are shown in Fig 1 and Fig2.
3. We also manually classified 400 sentences extracted from cleaned BSVs. We classified these sentences as bioagressor and disease if the BSV says the threshold of danger is reached, or if it recommends to apply a treatment. This classification task aims to test if the language model can “understand” the risk.

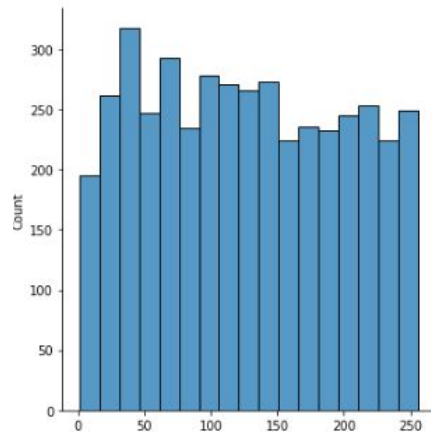


Fig 1. Length distribution of text chunks

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

Fig 2. Label counts by hazard

Results - Prediction of topic

To evaluate all these classifications, we use accuracy, precision, recall, F1 score and ROC_AUCscore Hossin and M.N (2015)

TAB. 1 – *prediction of the topic (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

TAB. 2 – *prediction of the topic (threshold=0.5) using BERT-Base, Multilingual Cased model*

	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

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

“... 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.”

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

Results - Prediction of risks

To evaluate all these classifications, we use accuracy, precision, recall, F1 score and ROC_AUCscore Hossin and M.N (2015)

TAB. 3 – *prediction of risks:*

	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

Considering the risk level or the detection of the positive/negative sense of the phrase, the prediction is less pertinent. For example, phrases like the following are still classified to having a risk of bioagressor even though it says there is only a small presence of bioagressor so no action is required. These results may be improved if more data is available

“... note l'apparition des premiers pucerons à villenauze 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.”

Conclusion

Our results show that fine-tuned BERT-based model is sufficient for the classification of BSV.

The preliminary prediction results convinced us that BERT-based models are capable of representing features in the French plant health domain.

Future Works:

- Feed the model with more and more pertinent data
- Explore alternatives such as FlauBERT Le et al. (2020)
- Investigate feature-based approaches with BERT embeddings