ChouBERT: Pre-training French Language Model for Crowdsensing with Tweets in Phytosanitary Context

Shufan Jiang, Rafael Angarita, Stéphane Cormier, Julien Orensanz, Francis Rousseaux

ChouBERT: Pre-training French Language Model for Crowdsensing with Tweets in Phytosanitary Context

Shufan Jiang[1,2][0000−0002−8486−3158], Rafael Angarita[1][0000−0002−2025−2489], Stéphane Cormier[2][0000−0003−4507−4815], Julien Orensanz[3], and Francis Rousseaux[2]

[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
[3] Cap2020, Gradignan, France

Abstract. To fulfil the increasing need for food of the growing population and face climate change, modern technologies have been applied to improve different farming processes. One important application scenario is to detect and measure natural hazards using sensors and data analysis techniques. Crowdsensing is a sensing paradigm that empowers ordinary people to contribute with data their sensor-enhanced mobile devices gather or generate. In this paper, we propose to use Twitter as an open crowdsensing platform for acquiring farmers knowledge. We proved this concept by applying pre-trained language models to detect individual's observation from tweets for pest monitoring.

Keywords: transfer learning · crowd-sensing · plant health monitoring · twitter

1 Introduction

Crowdsensing is a sensing paradigm that empowers ordinary people to contribute with data sensed from or generated by their sensor-enhanced mobile devices [1]. It introduces a new shift in the way we collect data by permitting to acquire local knowledge through smart devices carried by people, such as smartphones, tablets, smartwatches, among others. This allows to leverage 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 digitization of agriculture [3], we have witnessed the emergence of crowdsensing applications for smart farming [7]. Farmers are more than ever present in social media such as Facebook, WhatsApp, and Twitter [12], where they report their issues and discuss. They also search solutions to existing problems in online groups. Particularly, Twitter allows farmers to freely publish short messages called "tweets"

to share their observations. Taking advantage of these observations requires of keeping track of relevant data sources among 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.

Recent applications of large-scale pre-trained language models seem promising for tackling domain-specific information extraction problems from text in French [11,4]. In this work, we propose to build ChouBERT, a pre-trained language models that "learns" knowledge in plant health domain from French plant health bulletins (BSV, for *Bulletin de Santé du Végétal* in French) and recognizes similar syntax in tweets for detecting farmer's observations in French phytosanitary context. Driven by **the increasing connectivity of farmers and the emergence of online farming communities**, our goal is to explore the emerging application of on-farm 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?*; and *RQ2. how to further pre-train general LMs for domain specific text classification?*.

In the next section, we review related work. Then, we formalize the problem and our approach in Section 3. We present our solution ChouBERT in Section 4, we discuss the threats to validity in Section 5, and we give our conclusion in Section 6.

2 Related work

In regard to existing works on plant health monitoring using Twitter, [16] builds 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. [10] 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.

3 Approach

We define individuals' observation as: a description of the presence of pest in a field in real time. 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.

These tweets may be missing essential information, such as precise location, impacted crop, the current developing status of a pest and the estimation of upcoming damage. The knowledge that helps to recognize farmers' observations can be found in: vocabularies of French crop usage [9] as formal knowledge; BSV as semi-structured domain knowledge; pre-trained LMs as knowledge of French language; tweets labelled by domain experts, containing tactical knowledge; and unlabelled tweets concerning crops or plant health issues, as a corpus of the syntax of tweets.

Fig. 1 illustrates an overview of the different steps of our approach : i), data preparation: we collected tweets using keywords . 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; ii), further pre-training: adjust the weights of pre-trained encoders using Masked Language Model (MLM) [2] 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; and ii), supervised tweet classification: we train classifiers with different LM representations to distinguish observations from other information.

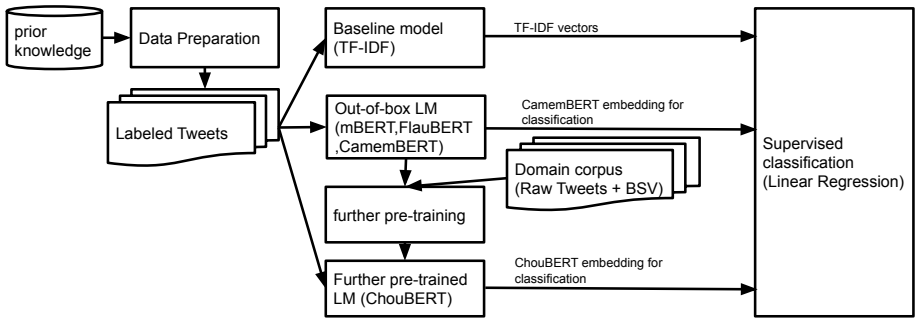

Fig. 1: Overview of the steps of the approach in the study

4 Experiments

4.1 Data preparation

We worked in collaboration with Arvalis (www.english.arvalisinstitutduvegetal.fr) to label tweets concerning observations about 5 different natural hazards. For each of these hazards, we invited plant health researchers to label tweets according to their judgement of their pertinence. Table 1 shows the composition of our labelled set. We collect tweets for at least 2 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 labelled 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 contains many unseen noises.

Table 1: Composition of the labelled set

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

We downloaded 40828 BSVs [14] from data.gov.fr. 17286 BSV are in XML, and 23542 are in plain-text. A first processing is to convert XML file to plain text. A cleaning step is also necessary to remove punctuation artifacts or out-of-context agricultural information such as phone numbers. To teach the LM the characteristic of tweets, we collected tweets containing terms in a list of 669 keyword concepts in the plant health domain between January 2015 and September 2021. We used insect pest names and plant diseases names in former PestObserver website [15], and the literal value of *skos:prefLabel* and *skos:altLabel* of all nodes having type skos:Concept in FrenchCropUsage thesaurus to construct the list.

4.2 Experimental setup

We conducted all experiments on a workstation having Intel Core i9-9900K CPU, 32 GB memory, 1 single NVIDIA GeForce RTX 3090 GPU with CUDA 10.0.130. We downloaded the LM from transformers [17]. We use fast-bert [13] wrapper for the further pre-training and the training of classifiers using linear regression. The choice of hyperparameters (see in Table 2)is based on the recommendation of BERT [2] and the configuration of our workstation. We did not do grid search for all hyperparameters on all the models for simplicity. For the further pre-training, we use implementation *CamembertForMaskedLM* in the transformer package (`https://huggingface.co/transformers/v3.0.2/`). We test different recipes to construct different corpus with the BSVs and tweets. We evaluate the further pre-trained LMs on the classification task.

Table 2: Hyperparameters for further pre-training and for classification

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

Due to the small size of our labelled data, we perform 5-fold cross-validation keeping the same separation for our labelled set. We used these 5 labelled sets for all the classification experiments, including the experiments with baseline model and the pre-trained LMs. Figure 2 shows the number of positive labels and negative labels in each fold of training/validation set. We use the following implementations in the transformer package: *BertForSequenceClassification*, *CamemBert-*

ForSequenceClassification and *FlauBertForSequenceClassifiaction* as classifiers, each of which is a linear layer on top of the pooled output of the LM.

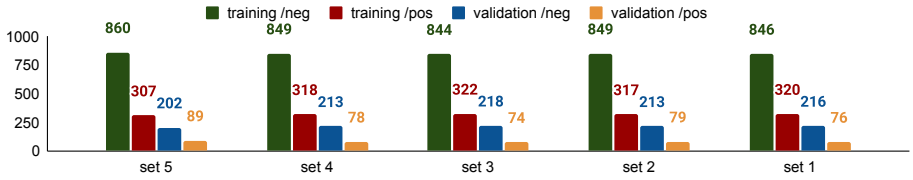

Fig. 2: Label distribution in each fold of training/validation set

To align with the linear classifiers in the transformer package, and to be differentiated from contextualized representations, we choose to fit the term frequency-inverse document frequency (TFIDF) vector of each tweets on linear regression classifier in sklearn package [8] 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 gives best average precision scores on the classification task, the average of which is 0.737186.

As presented above, 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. To have a general measure of performance irrespective of any particular threshold, we use average precision score in sklearn package [8], which estimate the area under the Precision-recall curve (AUCPR) 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.

4.3 Results and evaluation

Choosing the Out-of-box LM To find the most pertinent out-of-box LM for the further pre-training on our domain corpus, we perform the classification task with the embeddings given by the following LM: CamemBERT(camembert-base and camembert-large) [6], FlauBERT (flaubert-base-uncased and flaubert-large-cased) [5], and mBERT (bert-base-multilingual-uncased) [2]. For each LM, we note the score with best average precision scores in Table 3. All the models give better representation for classification than the baseline model (0.737186), which favour contextualized embeddings. CamemBERT models (denoted by CMB in the table) outperforms FlauBERT models (denoted by FLB). There are no significant differences between the base and large models. Thus, we choose to further pre-train Camembert-base with our corpus, and we use the classification results of CamemBERT models as our state-of-the-art models.

Table 3: Average precision scores of classification with out-of-box LMs

model	mBERT	CMB$_{large}$	CMB$_{base}$	FLB$_{base}$	FLB$large$
avg. APS	0.789853	**0.861968**	0.855935	0.845478	0.845027

The further pre-training There are two mainstream strategies to pre-train domain specific LMs: either further pre-train the weights of an existing model on domain specific corpus without touching its vocabulary and tokenizer like [4], or pre-train a new LM on domain specific corpus and a tokenizer from scratch like JuriBERT [11]. 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 those pre-trained from scratch. Thus, we decided to further pre-train existing LM on our corpus, reusing the native vocabularies and tokenizers. As we have too few tweets labelled as observation, we let BSVs teach the LM about context of observations and let unlabelled tweets to teach the LM the language style of tweets. 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, so it can learn the contextual information of the tokens, or how the tokens are organized together.

We feed the out-of-box CamemBERT base model with 3 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 (denoted by "CHB" in the table) on the classification task and note the best performance of each model in Table 4. We can see that all three ChouBERT models have better scores than CamemBERT models, ChouBERT$_{BSV+Tweets}$ has the best results. It seems that CamemBERT do 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 properly feed.

Table 4: Average precision scores of classification with further pre-trained LMs

model	CHB$_{Tweets}$	CHB$_{BSV}$	CHB$_{BSV+Tweets}$	CMB$_{large}$	CMB$_{base}$
avg. APS	0.874741	0.865134	**0.887424**	0.861968	0.855935

The generalizability on unseen hazards From the previous experiments, we selected the best hyperparameters (10 for epochs, 16 for batch size) for classification to study the effect of further-pretraining 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 3.

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

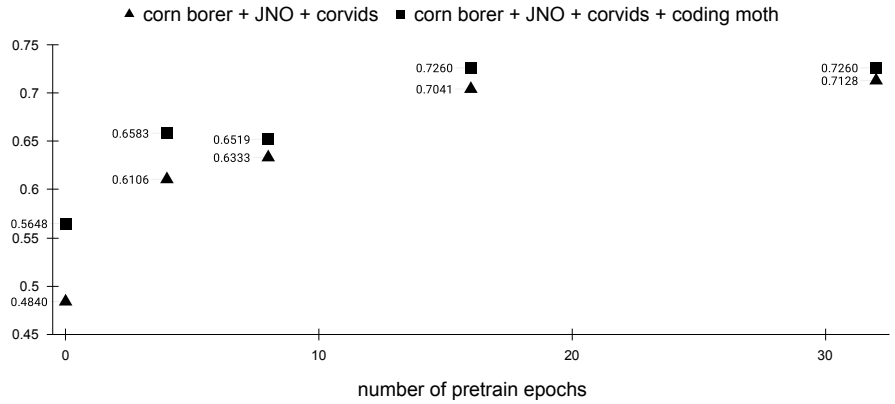

Fig. 3: Performance of different classifiers on wireworm tweets

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

5 Threats to validity

For the labelling of the tweets about observation, we did not take into account the observations concerning the absence of hazards. Neither did we rank the pertinence or the completeness of the observations, which should be considered in future work. 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.

6 Conclusion and future work

In this work, we presented a method to exploit crowd observations on Twitter. We built ChouBERT by applying domain adaptive pre-training to CamemBERT on BSV and tweets . 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 tweets using keywords; manually labelling a small set of tweets; further pre-training language models using domain documents and tweets; and building NLP applications with the labelled set and the domain-adapted LM. For future work, we plan to evaluate our model on other NLP tasks; to study for the integration of heterogeneous text and the building of a knowledge base ; 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 of other monitoring paradigms, but a complementary source of information to detect weak signals, rather than quantifying the gravity of an issue.

References

1. Boubiche, D.E., et al.: Mobile crowd sensing–taxonomy, applications, challenges, and solutions. Computers in Human Behavior **101**, 352–370 (2019)
2. Jacob, D., et al.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1. pp. 4171–4186. ACL (Jun 2019)
3. Klerkx, L., Jakku, E., Labarthe, P.: A review of social science on digital agriculture, smart farming and agriculture 4.0: New contributions and a future research agenda. NJAS - Wageningen Journal of Life Sciences **90-91**, 100315 (2019)
4. Laifa, A., Gautier, L., Cruz, C.: Impact of Textual Data Augmentation on Linguistic Pattern Extraction to Improve the Idiomaticity of Extractive Summaries. In: Big Data Analytics and Knowledge Discovery, vol. 12925, pp. 143–151. Springer, Cham (2021), series Title: Lecture Notes in Computer Science
5. Le, H., et al.: Flaubert: Unsupervised language model pre-training for french. In: Proc. of The 12th Language Resources and Evaluation Conference, LREC. pp. 2479–2490. European Language Resources Association, Marseille, France (2020)
6. Martin, L., et al.: Camembert: a tasty french language model. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 7203–7219. Association for Computational Linguistics, Online (2020)
7. Mendes, J., et al.: Smartphone applications targeting precision agriculture practices—a systematic review. Agronomy **10**(6), 855 (2020)
8. Pedregosa, F., et al.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
9. ROUSSEY, C.: French Crop Usage (2021), `https://doi.org/10.15454/QHFTMX`
10. Shankar, P., Bitter, C., Liwicki, M.: Digital Crop Health Monitoring by Analyzing Social Media Streams. In: 2020 IEEE / ITU International Conference on Artificial Intelligence for Good (AI4G). pp. 87–94. IEEE, Geneva, Switzerland (2020)
11. Stella, D., et al.: JuriBERT: A masked-language model adaptation for French legal text. In: Proc. of the Natural Legal Language Processing Workshop. pp. 95–101. Association for Computational Linguistics (Nov 2021)
12. Thollet, B.: MEDIAS SOCIAUX EN AGRICULTURE : Contribution à l'analyse des usages et de leur potentiel d'apprentissage pour la transition agroécologique. Master's thesis, AgroSup Dijon, Dijon, France (Aug 2020)
13. Trivedi, K.: Fast-bert. `https://github.com/kaushaltrivedi/fast-bert` (2020)
14. Turenne, N.: Reports ocr. `https://www.data.gouv.fr/fr/datasets/r/c745b0bf-b135-4dc0-ba04-1e15c1b77899` (2016)
15. Turenne, N., et al.: Open data platform for knowledge access in plant health domain : VESPA mining. CoRR **abs/1504.06077** (2015)
16. Welvaert, M.e.a.: Limits of use of social media for monitoring biosecurity events. PLOS ONE **12**(2) (2017)
17. Wolf, T., 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. pp. 38–45. Association for Computational Linguistics, Online (Oct 2020)