HAL
open science

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

Shufan Jiang, Rafael Angarita, Stephane Cormier, Francis Rousseaux

To cite this version:

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. hal-03122939

HAL Id: hal-03122939

https://hal.science/hal-03122939v1

Submitted on 28 Jan 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

Shufan Jiang*,**, Rafael Angarita*,
Stéphane Cormier**, Francis Rousseaux**

*Institut Supérieur d'Electronique de Paris, LISITE, Paris, France
name.lastname@isep.fr,
** Université de Reims Champagne Ardenne, CReSTIC, Reims, France
name.lastname@univ-reims.fr

Abstract. 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 precisely study natural hazards within global or local aspects, and then improve the risk prevention tasks and augment the yield, which helps to tackle the challenge of growing population and changing alimentary habits. In particular, French Plants Health Bulletins (BSV, for its name in French Bulletin de Santé du Végétal) give information about the development stages of phytosanitary risks in agricultural production. However, they are written in natural language, thus, machines and human cannot exploit them as efficiently as it could be. Natural language processing (NLP) technologies aim to automatically process and analyze large amounts of natural language data. Since the 2010s, with the increases in computational power and parallelization, representation learning and deep learning methods became widespread in NLP. Recent advancements Bidirectional Encoder Representations from Transformers (BERT) inspire us to rethink of knowledge representation and natural language understanding in plant health management domain. The goal in this work is to propose a BERT-based approach to automatically classify the BSV to make their data easily indexable. We sampled 200 BSV to finetune the pretrained BERT language models and classify them as pest or/and disease and we show preliminary results.

1 Introduction

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 precisely study natural hazards within global or local aspects, and then improve the risk prevention tasks and augment the yield, which helps to tackle the challenge of growing population and changing alimentary habits. In particular, French Plants Health Bulletins (BSV, for its name in French Bulletin de Santé du Végétal) give information about the development stages of phytosanitary risks in agricultural production Jiang et al. (2020). BSV are published periodically by the French Regional Plant Protection Services (SRPV) and Groupings Protection against

Harmful Organisms (GDON). These data is collected by the observation network Epiphyt [1] to build a national database. These data are collected following agronomic observations made throughout France by regional monitoring networks involving 400 observers in 36 partner networks.

Following the *social sensing* paradigm Wang et al. (2015), 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, less formal than the BSV but relevant and usually real-time hazard information is also published in social networks such as Twitter. Given the nature of such publications, it is not straightforward to efficiently and effectively take advantage of the information they contain, let alone doing it automatically and relating these data to data coming 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.

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 Hirschberg and Manning (2015). A well-known feature representation technique for text classification is Word2Vec Goldberg and Levy (2014) and some real-world applications of text classification are spam identification and fraud and bot detection Howard and Ruder (2018). PestObserver proposedto index BSV with crops, bioagressors and diseases based on concurrence analysis, regex and pattern matching text classification techniques Turenne et al. (2015). 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. This is an efficient and precise approach, but it cannot represent the latent semantic relations. Human annotation was also needed to enrich the project annotation resources and dictionaries. In overall, the global PestObserver approach relies on highly crowdsourcing-dependent techniques, which made the information extraction procedure not dynamic enough to adapt to changes in document format or contents.

Recent advancements in Bidirectional Encoder Representations from Transformers (BERT) Devlin et al. (2019) have showed important improvements in NLP; for example, the efficiency of fine-tuned BERT models for Multi-Label tweets classification has been proved in disaster monitoring Zahera (2019). We find it also important to include observation from non-expert individuals which are not necessarily in the network of experts collecting data. Information about such observation may be extracted from social media like twitter. Thus, our goal is to build a general language model that represents the phytosanitary risks to improve the information extraction and document classification from heterogeneous textual data sources. The BSV might serve as a corpus to train this language model.

In this work, we aim to explore the potentials 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 PestObeserver? And how well can it generalize for natural hazard prediction from heterogeneous documents?

1. https://agroedieurope.fr/wp-content/uploads/fiche-projet-epiphyt-fr.pdf

2 Data

Existing Dataset

A) We downloaded BSVs Turenne 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* 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.

C) We use concepts in FrenchCropUsuage thesaurus FAIRsharing Team (2018) and the tags in item B) as filters to collect tweets for testing the classification model.

Linguistic Prepossessing 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.

Dataset Construction

— 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.

— For the classification of the topic, we randomly split 200 cleaned BSVs into 4000 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)-.

— 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.

3 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, trasformersWolf et al. (2020) and fast-bertTrivedi (2020). For the multi-label classification of the topic, we tested respectively the pretrained CamemBERT modelMartin et al. (2020) and BERT-Base, Multilingual Cased model Pires et al. (2019). We first fine-tune the language model to adapt it to domain specific context. All hyper-parameters are tuned on raw BSV corpus over 2 epochs as suggested by the author Devlin et al. (2019). The batch size is 8. Adam Kingma and Ba (2017)is used for optimization with an initial learning rate of 1e-4. Then we used these fine-tuned language models to train the classification task. The batch size is 8. Max sequence length is set to 256. AdamKingma and Ba (2017)is used 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 better f1 score. The model's output are the probabilities of all classes, we hence set threshold value of 0.5 to pickup a list of possible classes to the input text as final prediction. Finally we evaluate the model over the test set, we also test the model with tweets that talks about natural hazards. For the multi-label classification of risk, the goal is to predict the presence of bioagressor risk or disease risk, the experiment setup is the same as the previous task, except that we fine-tuned the models on raw BSV corpus and on paragraphs extracted from BSV in XML format -see item A)-.

4 Results

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

Table 1 shows results of multi-label classification (bioagressor and disease) task. As the dataset is simply tagged with the appearance of key words, moreover, the tags on pestobserver site are not completed, the pertinence of its categorisation is limited. However, we observed that the model can correct some of false negative taggings from pestobserver (which means, a phrase which mentions borer and which 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.

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

Table 2 shows the results of the same multi-label classification task with BERT-Base, Multilingual Cased model. The scores are slightly better than the ones produced by CamemBERT presented in Table 1; however, the size of the pre-trained BERT-Base, Multilingual Cased model is bigger than CamemBERT -since it covers more than 104 languages- and it takes more time for the training.

Table 3 shows the results of multi-label classification of risks task. In this experiment, the pertinence of the training set is assured by manual annotation. We noticed that the language model is able to detect the presences of bioagressor or disease in the text though the dataset is much smaller than the one of previous classification task, which is equivalent to filter the document with a given domain key words list. 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 tough 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 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."

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

5 Conclusion

Recent advancements BERT are promising regarding natural language processing. Our objective is to classify agricultural-related documents according to natural hazards they discuss. We have studied existing textual data in french plant health domain, specially the BSVs, and experimented with the BERT multilingual and CamemBERT models. Our results show that fine-tuned BERT-based model is sufficient for the topic prediction of BSV. The preliminary prediction test on tweets convinced us that BERT-based models is generalizable for representing features in the French plant health domain. For our future work, we plan to feed our model with more pertinent data. It may be also interesting to explore alternatives such as FlauBERTLe et al. (2020), which another BERT-based language model for French. Finally, We also plan to investigate feature-based approaches with BERT embeddings.

References

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

FAIRsharing Team (2018). Fairsharing record for: French crop usage.

Goldberg, Y. and O. Levy (2014). word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*.

Hirschberg, J. and C. D. Manning (2015). Advances in natural language processing. *Science 349*(6245), 261–266.

Hossin, M. and S. M.N (2015). A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining Knowledge Management Process 5*, 01–11.

Howard, J. and S. Ruder (2018). Universal language model fine-tuning for text classification.

Jiang, S., R. Angarita, R. Chiky, S. Cormier, and F. Rousseaux (2020). Towards the Integration of Agricultural Data From Heterogeneous Sources: Perspectives for the French Agricultural Context Using Semantic Technologies. In *International Workshop on Information Systems Engineering for Smarter Life (ISESL)*, Grenoble, France.

Kingma, D. P. and J. Ba (2017). Adam: A method for stochastic optimization.

Le, H., L. Vial, J. Frej, V. Segonne, M. Coavoux, B. Lecouteux, A. Allauzen, B. CrabbÃ©, L. Besacier, and D. Schwab (2020). Flaubert: Unsupervised language model pre-training for french.

Martin, L., B. Muller, P. J. O. Suárez, Y. Dupont, L. Romary, E. de la Clergerie, D. Seddah, and B. Sagot (2020). Camembert: a tasty french language model. *ArXiv abs/1911.03894*.

Pires, T., E. Schlinger, and D. Garrette (2019). How multilingual is multilingual bert? *CoRR abs/1906.01502*.

Trivedi, K. (2020). Fast-bert. `https://github.com/kaushaltrivedi/fast-bert`.

Turenne, N. reportsocr.zip. `https://www.data.gouv.fr/fr/datasets/r/c745b0bf-b135-4dc0-ba04-1e15c1b77899`.

Turenne, N., M. Andro, R. CORBIÈRE, and T. T. Phan (2015). Open Data Platform for Knowledge Access in Plant Health Domain : VESPA Mining. working paper or preprint.

Wang, D., T. Abdelzaher, and L. Kaplan (2015). *Social sensing: building reliable systems on unreliable data*. Morgan Kaufmann.

Wolf, T., L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Online, pp. 38–45. Association for Computational Linguistics.

Zahera, H. M. (2019). Fine-tuned bert model for multi-label tweets classification. In *TREC*.