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

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

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

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. 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 urgent. The question is whether social media like Twitter can help farmers to share their observations towards the constitution of agricultural knowledge and monitoring tools. In this work, we develop several scenarios to collect tweets, then we applied different natural language processing techniques to measure their informativeness as a source for phytosanitary monitoring.

Keywords: Crowdsensing · Social media · Smart Agriculture · NLP.

1 Introduction

Facing the challenge of growing population and changing alimentary 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 [8]. Phytosanitary issues, including (a), abiotic stresses such as weeds, insect pests, animals, or pathogenic agents injurious to plants or plant products, and (b), biotic stresses such as floods, drought, extremes in temperature, can cause loss in food production. An important subject in precision agriculture is to improve the risk prevention tasks and measuring natural hazards within global and local aspects through real-time monitoring. We can classify mainstream real-time monitoring technologies of natural hazards into two categories [10]: (i), indirect monitoring by analysing environment parameters produced by sensor networks and Internet of Things (IoT) devices to infer the probability of phytosanitary risks [23];

and (ii), direct monitoring by processing images [25]. Current precision agriculture technologies favour large-scale monoculture practices that are unsustainable and economically risky for farmers [12]. Moreover, according to the Food and Agriculture Organization of the United Nations, farms of less than 2 hectares accounted for 84 percent of all farms worldwide in 2019, and most of these small farms are family farms [22].

We suggest that current observation data from precision agriculture cannot represent all forms of farms, especially small farms. Recently, how to encourage the participation of farmers to share their knowledge and observations is drawing the attention of researchers [17,16]. However, local observations of farmers are not taken sufficiently into account, which results in the loss of legitimacy and the vanishing of local traditional knowledge. As [14] points out, local farmer knowledge relies on social processes for knowledge exchange, but 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 [30].

The role of social media like Twitter in farmer-to-farmer and in farmer-to-rural-profession knowledge exchange is increasing, and 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/replies) dominating one-way messaging (new/ original tweets) [24]. Following the *social sensing* paradigm [32], 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 experience among them, to subscribe to topics of interest using hashtags and to share real-time information about natural hazards. Compared to paid applications, information on Twitter, presented in 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 su as #AgriChatUK (http://www.agrichatuk.org) or #FrAgTw (https://franceagritwittos.com), to their posts on Twitter [7]. Thus, we can consider Twitter as an open tool for farmer-to-farmer knowledge exchange. This paper tackles the following question: which phytosanitary information can be automatically extracted from textual contents on Twitter, and what is the quality of this information?

The rest of this paper is organized as follows: Section 2 introduces our use cases and the dataset we built; Section 3 presents the concordances between the popularity evolution of tweets and historical records of hazards; Section 4 explores tweet topics using unsupervised methods and the pretraining of language models; Section 5 resumes lessons learned and presents future work directions.

2 Use cases and data collection

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 (`https://www.cap2020.online/`) and Arvalis (`https://www.english.arvalisinstitutduvegetal.fr/index.jspz`), 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 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 [2]. 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 ("corvidé OR corbeau freux OR choucas de tour OR corneille" in French) 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 challenge 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 [6] and K-Means [28] 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 labelled tweets, we build a classifier for event detection.

We use the Twitter API to collect tweets. When using the API, the matching of keyword is applied to not only the text field of a tweet, but also the username of the author or the text content of the referenced tweet. Moreover, accented and special characters are normalized to standard latin characters, which can change meanings in foreign languages or return unexpected results. For example, "maïs", which is *corn* in French, will also match "mais" which means "but" in English. Thus, we isolated accented keywords as special cases: for each accented word, we pulled all the normalized-word-filtered tweets from the Twitter API, and then we filtered them again with the accented word in our database. We saved original tweets as well as re-tweets. We have collected in total 16345 tweets about"corn borer", 3302 tweets about "BYDV", 50902 tweets about straw cereals and 38903 tweets about "corvids".

3 Histogram by mention of keywords

Use case: corn borer First, we want to confirm that tweets do talk about corn borers. We used "pyral" as keyword to retrieve tweets from Twitter API, then we kept the tweets that contain "maïs" to construct the dataset. We plot the number of tweets by month and by year in Figure 1, and compare it with records of average corn borer number by trap from Arvalis (see Figure 2). In both figures, we can observe peaks of corn borer between May and August. There is an exception in Figure 1 since there are minor peaks in February, which correspond to the Paris International Agricultural Show (`https://en.salon-agriculture.com/`) when there people discussed about technologies to fight corn borers. Such exception shows that tweets collected by keywords is not precise enough.

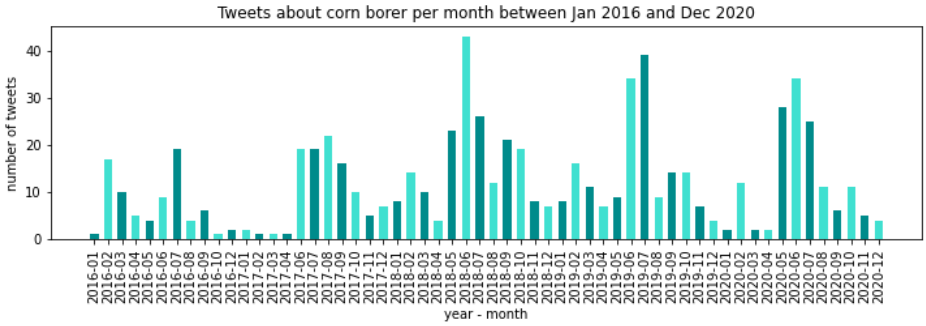

Fig. 1: Number of tweets containing "pyrale" and "maïs" by month between 2016 and 2020

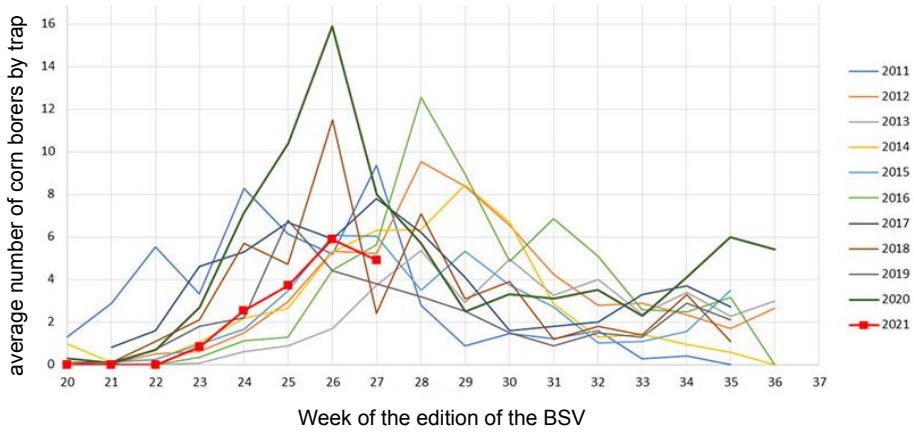

Fig. 2: Recorded averaged corn borer number by trap from [1]

Use case: yield of cereals We repeated the same data collection process for the yield of wheat. In this case, we used all the cereals in the French Crop Usage Thesaurus [27] to collect tweets (*céreales à pailles* in French). As these words are quite frequent, we add conditions to retrieve tweets containing "récolte", "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. 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). 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 [5]. 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 4. 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.

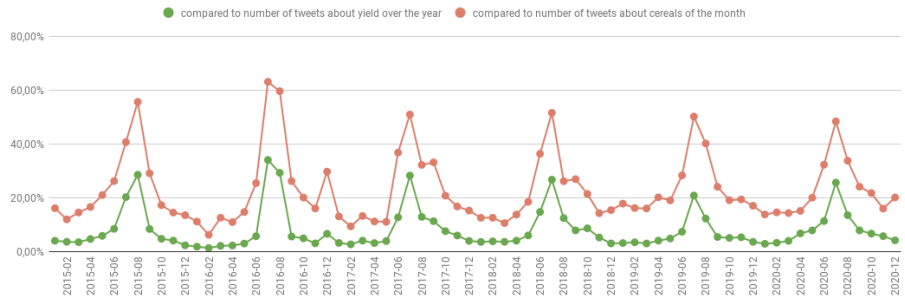

Fig. 3: Percentage of tweets concerning cereal yield between 2015 and 2020

Fig. 4: Counts of tweets mentioning yield and 2016

4 Processing tweets for natural hazard detection

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 see this task as an unspecified topic detection task [3]. Survey on topic detection [13] 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) [6]. These algorithms have been created to automatically divide a collection of data into groups of similarity for browsing, in a hierarchical or partitional manner [29]. Most of the measures of similarity, such as Euclidean distance or cosine distance, can be

only applied on data points in a vectorial space [15]. 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 occur in a collection of documents. In a BoW model, a document is a bag of words. Short for term frequency–inverse document frequency, TF*IDF is a formal measure of how important a term is to a document in a collection [26].

TF*IDF is defined as follows. Given a collection of N documents, define f_{ij} as the frequency of term (a word)t_i in document d_j. Then, define the term frequency TF_{ij} of term t_i in 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: $TF_{ij} = \frac{f_{ij}}{max_k f_{kj}}$ The IDF of a term describes how much information the term provides. Suppose term t_i appears in n_i documents. Then $IDF_i = log_2(N/n_i)$. The TF*IDF score for term t_i in document d_j is defined to be $TF_{ij} \times IDF_i$. The terms with highest TF*IDF scores are often the most relevant terms to represent the topic of the document. TF matrix and TF*IDF matrix are widely used for describing the features of the document [4].

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 [31]. This list also includes the keyword "BYDV", which brings also tweets in English. Therefore, we need to look into the topics in the 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 how many topics there are, 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 TF*IDF 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 TF*IDF vectors to K-Means [29], for K between 2 and 20, find the best cluster number K using elbow method [19].
4. Calculation of the TD*IDF matrix for each cluster, examination of the 20 terms with the highest TD*IDF scores, and 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 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 1: Top TF*IDF scored words in clusters in final state of K-Means based cleaning.

cluster	top TF*IDF 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, ceuxquifontlessessais, 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

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 [11], 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 [20]. Recent advancements in Bidirectional Encoder Representations from Transformers (BERT) [9] 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 [21], 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 [16] 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 labelled 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 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 [9]. 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 [21] to encode tweets and train the classifier. We set the max sequence length to 128 and batch size to 16. We use Adam [18] 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 2 shows the results of 5-fold cross validation of sigmoid classifier based on TDIDF vectors, and Table 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.

Table 2: Classification based on TF*IDF, 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: 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

5 Conclusion

In this paper, 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 potential to capture their contextual information, which can reduce manual labelling 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 analysing 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.

6 Ackownledgement

Thanks to Doriane HAMERNIG, Emmanuelle GOURDAIN, Olivier DEUDON, Jean-Baptiste THIBORD, Christophe GIGOT, François PIRAUX and Stéphane JEZEQUEL for their wise comments, their suggestions on the application scenarios and their contribution to Tweet annotation.

References

1. ARVALIS: Figure 2 : Evolution du nombre moyen de pyrale par piège selon l'année, https://www.arvalis-infos.fr/_plugins/WMS_BO_Gallery/page/getElementStream.jspz?id=72073&prop=image
2. ARVALIS: Jaunisse Nanisante de l'Orge (JNO) - Maladie virale sur Blé tendre, blé dur, triticale (2013), http://www.fiches.arvalis-infos.fr/fiche_accident/fiches_accidents.php?mode=fa&type_cul=1&type_acc=7&id_acc=53
3. Asgari-Chenaghlu, M.e.a.: Topic Detection and Tracking Techniques on Twitter: A Systematic Review. Complexity **2021**, 1–15 (Jun 2021)
4. Bafna, P., Pramod, D., Vaidya, A.: Document clustering: TF-IDF approach. In: 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT). pp. 61–66. IEEE, Chennai, India (Mar 2016)
5. Ben-Ari, T.e.a.: Causes and implications of the unforeseen 2016 extreme yield loss in the breadbasket of France. Nature Communications **9**(1), 1627 (Dec 2018)
6. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of Machine Learning Research **3**, 993–1022 (2003)
7. Defour, T.: EIP-AGRI Brochure Agricultural Knowledge and Innovation Systems (Feb 2018), https://ec.europa.eu/eip/agriculture/en/publications/eip-agri-brochure-agricultural-knowledge-and
8. Desa, U., et al.: Transforming our world: The 2030 agenda for sustainable development (2016)
9. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (2019)
10. Gao, D.e.a.: A Framework for Agricultural Pest and Disease Monitoring Based on Internet-of-Things and Unmanned Aerial Vehicles. Sensors **20**, 1487 (2020)
11. Goldberg, Y., Levy, O.: word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method. arXiv preprint arXiv:1402.3722 (2014)
12. Heldreth, C., Akrong, D., Holbrook, J., Su, N.M.: What does AI mean for smallholder farmers?: a proposal for farmer-centered AI research. Interactions **28**(4), 56–60 (Jul 2021)
13. Ibrahim, R., Elbagoury, A., Kamel, M.S., Karray, F.: Tools and approaches for topic detection from Twitter streams: survey. Knowledge and Information Systems **54**(3), 511–539 (Mar 2018)
14. Ingram, J.: Farmer-Scientist Knowledge Exchange. In: Encyclopedia of Food and Agricultural Ethics, pp. 1–8. Springer Netherlands, Dordrecht (2014)
15. Irani, J., Pise, N., Phatak, M.: Clustering techniques and the similarity measures used in clustering: A survey. International journal of computer applications **134**(7), 9–14 (2016), publisher: Foundation of Computer Science
16. Jiménez, D.e.a.: From Observation to Information: Data-Driven Understanding of on Farm Yield Variation. PLOS ONE **11**(3), e0150015 (Mar 2016)

17. Kenny, U., Regan, A.: Co-designing a smartphone app for and with farmers: Empathising with end-users' values and needs. Journal of Rural Studies **82**, 148–160 (Feb 2021)
18. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization (2017)
19. Kodinariya, T., Makwana, P.: Review on determining of cluster in k-means clustering. International Journal of Advance Research in Computer Science and Management Studies **1**, 90–95 (01 2013)
20. Kowsari, K.a.a.: Text Classification Algorithms: A Survey. Information **10**(4), 150 (Apr 2019), arXiv: 1904.08067
21. Louis, M.e.a.: Camembert: a tasty french language model. ArXiv **abs/1911.03894** (2020)
22. Lowder, S., Sánchez, M., Bertini, R., et al.: Farms, family farms, farmland distribution and farm labour: what do we know today? FAO Agricultural Development Economics Working Paper (2019)
23. Olatinwo, R., Hoogenboom, G.: Chapter 4 - Weather-based Pest Forecasting for Efficient Crop Protection. In: Abrol, D.P. (ed.) Integrated Pest Management, pp. 59–78. Academic Press, San Diego (2014)
24. Phillips, T., Klerkx, L., McEntee, M., et al.: 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. pp. 1–20. International Farming Systems Association (IFSA) Europe (2018)
25. Qing, Z.e.a.: A pest sexual attraction monitoring system based on IoT and image processing. Journal of Physics: Conference Series **2005**(1), 012050 (Aug 2021)
26. Rajaraman, A., Ullman, J.D.: Data Mining, p. 1–17. Cambridge University Press (2011). https://doi.org/10.1017/CBO9781139058452.002
27. ROUSSEY, C.: French Crop Usage (2021). https://doi.org/10.15454/QHFTMX, `https://doi.org/10.15454/QHFTMX`
28. Singh, V.K., Tiwari, N., Garg, S.: Document Clustering Using K-Means, Heuristic K-Means and Fuzzy C-Means. In: 2011 Int. Conference on Computational Intelligence and Communication Networks. pp. 297–301. IEEE, Gwalior, India (Oct 2011)
29. Steinbach, M., Karypis, G., Kumar, V.: A Comparison of Document Clustering Techniques (May 2000), `http://conservancy.umn.edu/handle/11299/215421`, accessed: 2021-12-13
30. Thareau, B., Daniel, K.: Le numérique accompagne les mutations économiques et sociales de l'agriculture. Sciences Eaux & Territoires **Numéro 29**(3), 44 (2019)
31. Turenne, N., Andro, M.: Maladies des cultures (Feb 2017). https://doi.org/10.5281/zenodo.268301
32. Wang, D., Abdelzaher, T., Kaplan, L.: Social sensing: building reliable systems on unreliable data. Morgan Kaufmann (2015)