Shufan JIANG

I’m a Temporary Lecturer and Research Staff (ATER) at Ecole Normale Supérieure, where I do research in VALDA Team, and participate in the teaching of the CERES.

I obtained a Ph.D. in Computer Science from the Reims Champagne-Ardenne University

#DigitAg Partner researcher

E-Mail : sufianj415@gmail.com
GitHub : https://github.com/sufianj

Contents

Research

My research focuses on natural language processing for knowledge reconstruction from textual and domain-specific data. During my thesis, I had the opportunity to collaborate with farmers, researchers, and experts in agronomy and in artificial intelligence. My multidisciplinary research work leads me to collaborate with computer scientists and non-computer scientists.

ATER

Since December 2022, I continue my research with pre-trained language models by contributing to the TheoremKB project within the Valda team at ENS. This is a research project and a collection of tools to reconstruct knowledge from (mathematical) research articles. A first phase is to recognize and retrieve theorems. Previous work validates that the application of deep learning and multimodal learning to model text formatting and semantics improves the identification of blocks of theorems in scientific articles in PDF format. A second phase is to extract fine-grained information. I am working on extracting structures (e.g. conclusion, hypothesis, reasoning by recurrence, etc.) from theorems.

PhD Thesis

Integrating textual data towards crowdsensing natural hazards in agriculture
Co-supervisors: Rafael Angarita, Stéphane Cormier, Raja Chiky, Francis Rousseaux
ISEP Paris and Reims Champagne-Ardenne University • CRESTIC • Paris and Reims • Oct 2019 -Dec 2022

Abstract

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

Keywords: `Artificial Intelligence` `Social Media` `Machine Learning` `Natural Language Processing` `BERTology` `Plant Health` `Knowledge Management`

Technologies/Tools : `Transformers` `BERT` `PyTorch` `Scikit-learn` `MongoDB` `TwitterAPI` `Python` `Conda` `RDF` `SKOS` `JupterLAB` `Colab`

Publications

1. Shufan Jiang, Stéphane Cormier, Rafael Angarita, Francis Rousseaux. Improving text mining in plant health domain with GAN and/or pre-trained language model. Frontiers in Artificial Intelligence, 2023, 6, Shufan Jiang, Stéphane Cormier, Rafael Angarita, Francis Rousseaux. **Improving text mining in**

plant health domain with GAN and/or pre-trained language model. *Frontiers in Artificial Intelligence*, 2023, 6, ⟨10.3389/frai.2023.1072329⟩. ⟨hal-04008864⟩

2. Shufan Jiang, Rafael Angarita, Raja Chiky, Stephane Cormier, Francis Rousseaux. **Towards the Integration of Agricultural Data From Heterogeneous Sources: Perspectives for the French Agricultural Context Using Semantic Technologies**. *Advanced Information Systems Engineering Workshops (CAiSE)*, 2020, Grenoble, France. ⟨hal-02536389⟩
3. 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⟩
4. Shufan Jiang, Rafael Angarita, Stéphane Cormier, Julien Orensanz, Francis Rousseaux. **Informativeness In Twitter Textual Contents For Farmer-centric Plant Health Monitoring**. *ICPRAI - 3rd International Conference on Pattern Recognition and Artificial Intelligence*, Jun 2022, Paris, France. ⟨hal-03615884v2⟩
5. Shufan Jiang, Rafael Angarita, Stéphane Cormier, Julien Orensanz, Francis Rousseaux. **ChouBERT: Pre-training French Language Model for Crowdsensing with Tweets in Phytosanitary Context**. *International Conference on Research Challenges in Information Science (RCIS)*, 2022, Barcelona, Spain. ⟨hal-03621123⟩
6. Shufan Jiang, Rafael Angarita, Stéphane Cormier, Francis Rousseaux. **Named Entity Recognition For Monitoring Plant Health Threats in Tweets: A ChouBERT Approach**. Accepted by IEEE UV 2022 - The 6th International Conference on Universal Village, 2022, Boston, USA.
7. Shufan Jiang, Stéphane Cormier, Rafael Angarita, Francis Rousseaux. **Improving text mining in plant health domain with GAN and/or pre-trained language model**. Frontiers in Artificial Intelligence, 2023, 6, ⟨1072329. 10.3389/frai.2023.1072329⟩.
8. Shufan Jiang, Rafael Angarita, Raja Chiky, Stephane Cormier, Francis Rousseaux. **Vers la Reconstruction des Connaissances Agricoles : Perspectives de Détection des Risques Naturels à partir de Sources de Données Hétérogènes**. *Extraction et Gestion des Connaissances (EGC)*, 2021, Montpellier, France. 2021. ⟨hal-03066102⟩ (Poster)
9. Shufan Jiang, Rafael Angarita, Stéphane Cormier, Julien Orensanz, Francis Rousseaux. **Informativité dans les Contenus Textuels Twitter pour la Phytosurveillance Centrée sur l'observation des Agriculteurs**. *Extraction et Gestion des Connaissances (EGC)*, Jan 2022, Blois, France. ⟨hal-03480716⟩ (Poster)

Presentations

- Several participations in seminars at LISITE and at CReSTIC, to present research work.
- Talk at the "Initialisation à la recherche" course at ISEP to get new research ideas and to motivate more students to work in research.
- Two participations in the Francophone Conference on Knowledge Extraction and Management (EGC), in 2021 and 2022, to present posters

- Participation in international conferences (RCIS2022, ICPRAI2022, IEEE UV 2022, ISESL2020 workshop linked to CAiSE2020) to present accepted papers.
- Workshop "Agricultural risk detection from Twitter", invited by the 3rd Thematic School on Microservices and Big Data Management, to demonstrate the methodologies and results of the PhD, 26-27 January 2023 at Paris Nanterre University.

Contributions to tools and corpora

- ChouBERT (https://huggingface.co/ChouBERT) is a language model that integrates knowledge in plant health bulletins and tweets.
    - ChouBERT-n: language models pre-trained on tweets and French plant health bulletins to encode French texts that improve information extraction tasks in the plant health domain. ChouBERT captures the context well to handle polysemies like "chou".
    - ChouBERT-n-plant-health-tweet-classifier to classify whether a tweet is about a plant health observation.
    - ChouBERT-n-plant-health-ner to annotate pests and diseases of crops.
- French Tweets about Plant Health (https://zenodo.org/record/5853684): A collection of French tweets about plant health.

Other Research Project

Comparison of opinions expressed in Twitter with ANES aggregated data

SICSS 2022 RWTH Aachen & TU Graz • research project • Aug 2022
Description: The American National Election Studies (ANES) are national representative surveys of American eligible voters that have been conducted before and after every presidential election since 1948. The focus of the survey includes voter perceptions of the major political parties, the candidates, national and international issues.

Here we asked whether opinions expressed by Twitter users match, in an aggregated manner, those gathered in the ANES (after controlling for demographics) or if it is true that extreme opinions are overrepresented and more salient in social networks
Research questions and hypotheses:

- As seen in ANES data, negative feelings towards the opposite party are increasing (Iyengar et al., 2019). Could we observe this trend using Twitter data?
- Just before every American election, there is an important and controversial event that captures media attention (aka “October Surprise”). How do these events impact Twitter? Are there more tweets related to the election? Are there more negative tweets? Do the negative tweets affect both candidates, or only the one involved in the scandal?

Keywords:`ANES` `Sentiment Analysis` `Social Media`

Codes and slides: https://github.com/sufianj/sicss2022_aachen-graz_idea13

Personalized Healthy Meal Recommendation

Kiolos • France • Master Final year research • Sep 2014 – Jan 2015

Subject: The objective of this project is to propose a distributed implementation of SVD for Kiolos Meal Recommendation application. Our work was as follows:

- Studies on recommendations mythologies: Collaborating Filtering, Knowledge base
- Studies on the algorithms for machine learning: Stochastic Gradient Descent (SGD), Alternating Least square (ALS), Singular Value Decomposition (SVD)

Keywords:`Recommender System` `SVD` `MapReduce` `Hadoop` `Mahout`

Professional Experiences

Temporary Lecturer and Research Staff (ATER)

Ecole Normale Supérieure • Computer Science Department • Paris • Since 01/12/2022

- **Administrative and organizational tasks** for the competitive entrance examination
- **Teaching** :
  - “The well-being of farmers under the agroecological transition” is a series of original workshops (equivalent to 24 hours of lectures / guided lab) co-hosted by researchers from different disciplines in the education center about the environment and the society. We propose to work on the factors of well-being (ill-being) associated with this change in the agricultural system. We first discuss different possible indicators, and then the students will choose a qualitative (survey) or quantitative (data mining) method to study these indicators and better understand these transitions. My role is to share my know-how about text mining, agricultural knowledge management and social crowdsensing with students and to provide technical coaching for their quantitative research.
  - I also create a course "Words for environmental ills: text mining about the Conferences of the Parties (COP)". This opening course (equivalent to 12 hours of lectures + 12 hours of programming in Python) aims to introduce different technologies in text mining (natural language processing, deep learning), network analysis and existing semantic resources to study various socio-political topics related to the environment in barely exploited textual data.
- **Research**: participate in the TheoremKB project, see in previous session.

Doctoral contract

Reims Champagne-Ardenne University: 27/11/2019 – 26/11/2022

Funding : 50% Reims Champagne-Ardenne University, 50% ISEP Paris

Research : see in previous session.

Teaching :

- Teaching assistant: Database Lab — Reims Champagne-Ardenne University, Dec 2019
- Lecture:
  - Introduction to Ontologies and Semantic Web — Reims Champagne-Ardenne University, Dec 2020
  - Invited talks about AI and Smart Agriculture, as part of the “Artificial Intelligence and Machine Learning” master course — Galatasaray University, 2020, 2021 and 2022
- Co-advising Master students:

Year	Name	Project	
2020 - 2021	Gillian GUEGUEN	Knowledge Management for agricultural data	Internship, ISEP
2020 - 2021	Matteo DAVID	CNN for text classification: Natural language processing for Plant Health Bulletins	Research project, URCA
2021 - 2022	Jeremy JOUBE	Topic Modeling with Tweets for Pest Monitoring	Research project, URCA
	Nouhaila LASIRI	Data augmentation for tweet classification	
	Dylan BAPTISTE	Evaluation of the contribution of an antagonistic neural network for a semi-supervised classification of French texts from a poorly labelled dataset	

Junior Technical Architect
BNP Paribas Securities Services • IT Architecture Team • Pantin • 01/02/ 2015 – 30/09/ 2019

Chatbot ---- Feb 2015 – Mar 2018

`Stanford NLP` `WordNet` `LanguageTool` `Alix` `AIML` `Java` `Spring` `Hibernate` `Jenkins` `SonarQube` `PostgreSQL` `PL/SQL` `bash/ksh` `Angular` `Extjs`

2STalk is a chatbot that guides users to good documentation or to the tool among numerous available resources. It is also a collaborative platform for knowledge management between the collaborators, according to their own profiles. The knowledge could be question/answer in plain text, or existing URLs, or real-time data from APIs. This chatbot consists of a graphic web UI and an API permitting its integration in the messenger of enterprise

- Studies and implementation of a chatbot based on AIML (Artificial Intelligence Markup Language) motor.
- Studies and prototyping a bilingual knowledge base (with multilingual extension) for the chatbot
- Semantic analysis in a technical context and in business lines of the bank.
- Realization of an NLP pipeline to extract keywords and intentions using WordNet, Stanford NLP, Alix and LanguageTool
- Conception, architecture, and development of the conversational agent and the platform for feeding the knowledge base

Prediction of abnormalities in the Trades ---- Apr 2018 – Jan 2019

`Random Forest` `Neural Networks` `Python` `Sklearn` `Pandas` `Numpy` `SparkMllib` `h2o` `weka` `Java` `Angular`

- Technological watch on the stack Java (weka, h2o, Spark ML) for industrialization of a Proof-Of-Concept (POC) constructed by data scientists, written in python scikit learn and nodejs
- Studies about the classification algorithms, particularly Random Forest and Neural Networks
- Reconstruction of the POC in Java/angular
- Collaboration with developers' team in Lisbon and in Chennai

Development and maintenance: monitoring Enterprise Application Integration (EAI) ---- Feb 2018 - Sep 2019

`Java` `Swagger(OpenAPI)` `Spring` `Hibernate` `Jenkins` `Oracle 11g` `PL/SQL` `bash/ksh` `Angular` `Extjs`

- Migration of current Continuous Integration (CI) plugins
- Test and analyse data exchange monitoring API

Reviewer for international journals / conferences

- Wiley, Concurrency and Computation: Practice and Experience (CCPE), November 2020
- Technology and Environment Workshop'21 at the Extraction et Gestion des Connaissances (EGC 2021) conference, Montpellier, France, January 2021.
- The 4th International Conference on Physics, Mathematics and Statistics (ICPMS2021), Kunming, Chine, May 2021

Personal Projects

Wolf-Api: Parser for the synonym dictionary WOLF (WordNet in French)

`WOLF` `Java` `xml` `Open Source`

Link to the project: https://github.com/sufianj/WolfApi

E-mail Classification

Crédit Agricole x IBM • Paris • Hackathon • final 10 Project • 3 – 5 june 2016

`Nodejs` `Neo4j` `IBM Watson` `IBM Bluemix`

Design and Realization of a working prototype for sorting and extracting the important information in unread e-mails, as back-end developer and Neo4j administrator

Education

Diplomas:

- PhD in Computer Science | ISEP Paris & Reims Champagne-Ardenne University | 2019 -2022
- Master in Architecture of Information Systems | ISEP Paris | 2013 - 2015
- Bachelor in Computer Science and Technology | Nanjing University of Aeronautics and Astronautics | 2010 - 2014

Other certificates:

- Course (online): Machine Learning | Coursera | 2018
- Course (online): Web sémantique et Web de données | FUN-MOOC | 2020
- Course (online): Business Intelligence for data mining and analysis in the context of agriculture and the environment | AgroParisTech - Montpellier & INRAE - Clermont-Ferrand | 2021
- Course (online): Text mining for agricultural and sanitary risks | ACTA | 2021
- Summer school on AI for Industry 4.0 | Mines Saint-Étienne | 2021
    - Hackathon: `Apache Jena` `Protégé Fuseki` `JaCaMo` `ThingDescription`
- Oxford Machine Learning Summer School | 2021
- Summer Institute for Computational Social Science | RWTH Aachen & TU Graz | 2022

Languages

```
French: fluent
English: fluent
Chinese (Mandarin and Cantonese): native
```

Hobbies

`Drawing` `Recycling & Recreation` `Hiking` `Culinary Arts` `Micro-Gardening`