

Baptiste Blouin

Ingénieur de recherche / NLP, machine learning

Aix-en-Provence, France
✉ BaptisteBlouin@proton.me
🌐 baptistebloin.fr
in baptiste-blouin
🔗 BaptisteBlouin
Français, anglais

Résumé professionnel

Ingénieur de recherche et docteur en informatique (NLP), auteur de 7 publications à comité de lecture (LREC-COLING, PACLIC, NLP4DH, TALN, JDMDH) et de conférences et ateliers dans 7 pays. Financé par la subvention ERC avancée ENP-China, j'ai conduit ma recherche en autonomie, de la formulation du problème à la publication, et dirigé techniquement HistText, la plateforme d'analyse de textes historiques du projet, déployée pour la communauté internationale de recherche. Entraînement de modèles sur les clusters HPC du LIS, du CC-IN2P3 et de Jean-Zay, construction de corpus annotés, transmission par ateliers. Depuis 2025, j'applique cette rigueur à des systèmes LLM en production.

Projets clés

HistText : plateforme d'analyse de textes à grande échelle

2023–2025

Recherche & développement full-stack

Publié : JDMDH 2024

- **Développé une plateforme d'analyse de textes** interrogeant plus de 15 millions de documents historiques multilingues, soit plus de 20 milliards de mots
- **Conçu une architecture évolutive** (backend Rust, Apache Solr) : recherche, concordances, NER et segmentation appliqués à des corpus entiers
- **Implémenté une solution ML full-stack** avec analyses en temps réel et visualisations interactives
- **Déployé la plateforme** auprès de plus de 200 chercheurs de la communauté internationale en humanités numériques
- **Stack technique** : Python, React, Rust, R, PostgreSQL, Apache Solr, Docker

Dataset NER chinois & pipeline ML

2023–2024

Ingénierie de données & ML

Publié : LREC-COLING 2024

- **Co-construit le plus grand dataset NER** de presse chinoise historique annoté à la main (1872-1949) : 196 documents, 6 491 entités, liage d'entités et coréférence
- **Développé des systèmes de contrôle qualité** automatisés pour datasets multilingues
- **Créé un pipeline d'entraînement ML** avec méthodologie d'évaluation
- **Stack technique** : Python, frameworks d'annotation, PostgreSQL, pipelines de validation

Modèle de langue chinoise & système de tokenisation

2023

Développement de modèles ML

Publié : NLP4DH 2023

- **Analysé les motifs linguistiques** du chinois transitionnel par investigation systématique
- **Développé des modèles de tokenisation** atteignant plus de 83 % de F-score, +35 points face à Jieba
- **Collaboré avec les chercheurs** d'Academia Sinica sur ce projet international
- **Stack technique** : PyTorch, TensorFlow, algorithmes de tokenisation, fine-tuning

Simulation d'erreurs OCR & robustesse de modèles

2022

Recherche expérimentale

Publié : TALN 2022

- **Étudié la performance des modèles ML** sous conditions de données bruitées
- **Développé une approche d'augmentation** réduisant l'impact d'erreur de 50 %
- **Évalué la robustesse des modèles** avec un framework de benchmarking
- **Stack technique** : PyTorch, Transformers, algorithmes de simulation, frameworks d'évaluation

Transfer learning pour l'adaptation de domaine

2021

Développement & optimisation ML

Publié : NLP4DH 2021

- **Appliqué le few-shot learning** atteignant 93 % de récupération de performance avec données minimales
- **Implémenté des architectures neuronales** character-aware pour traitement de texte en production
- **Évalué les modèles** sur plusieurs datasets avec analyse statistique
- **Stack technique** : BERT, transformers character-aware, techniques d'adaptation

Expériences professionnelles

Ingénieur IA & full-stack, Aubier (indépendant)

Sept. 2025 – présent

Plateforme SaaS IA de bout en bout : architecture, backend async, sécurité multi-tenant, frontend React.

Data scientist & ingénieur ML, Protisvalor, projet ENP-China, Aix-Marseille Université

Janv. 2023 – août 2025

- **Responsable technique et développeur principal** de la plateforme HistText, gestion de projet indépendante ; contribution à l'obtention d'une bourse ERC Proof of Concept pour sa valorisation
- **Exploration et résolution** de défis ML complexes dans le traitement de textes historiques à grande échelle
- **Communication de concepts techniques** en ateliers et conférences (NLP4DH 2023, LREC-COLING 2024)
- **Collaboration interdisciplinaire** avec historiens, linguistes et informaticiens internationaux
- **Entraînement distribué** de modèles de langue sur le supercalculateur Jean-Zay (IDRIS/GENCI), SLURM

Doctorant & scientifique ML, Protisvalor, LIS / IrAsia, Aix-Marseille Université

Sept. 2019 – déc. 2022

- **Recherche indépendante** sur subvention ERC avancée européenne, apprentissage auto-dirigé
- **Publié 4 articles à comité de lecture pendant la thèse (7 au total)**, rédaction technique et communication
- **Présenté la recherche** en conférences internationales (PACLIC, NLP4DH, TALN), prise de parole publique
- **Mentoré et collaboré** avec équipes de recherche internationales, leadership interculturel
- **Calcul HPC sur Jean-Zay** (IDRIS/GENCI) : entraînement de modèles NLP (NER, segmentation, transfert)

Compétences

NLP	NER multilingue, Extraction d'événements, Transfer learning, Adaptation de domaine, Segmentation de mots, Annotation de données, OCR, Traitement de textes historiques, spaCy
ML	PyTorch, TensorFlow, Hugging Face, scikit-learn, Sentence Transformers, NetworkX, UMAP, Analyse statistique, RAG, Agents IA, Évaluation LLM
Recherche	Méthodologie scientifique, conception expérimentale, revue de littérature, rédaction technique, présentations internationales, mentorat, collaboration interculturelle, LaTeX
Calcul	Jean-Zay / IDRIS, CC-IN2P3, Cluster du LIS, SLURM, Linux, Docker
Ingénierie	Python, Rust, R, SQL, React, PostgreSQL, Apache Solr, Elasticsearch, FastAPI, Git, CI/CD

Formations

Doctorat en informatique (TAL / machine learning), Aix-Marseille Université

2022

Thèse : Extraction d'événements à partir de fac-similés de documents anciens pour les études en histoire

Compétences de recherche : Investigation indépendante, test d'hypothèses, conception expérimentale, rédaction technique

Directeur : Prof. Benoît Favre | **Projet :** Subvention ERC avancée ENP-China (n° 788476)

Master Informatique (IA/ML), Aix-Marseille Université

2019

Licence Informatique, Aix-Marseille Université

2016

Publications

HistText: An Application for Leveraging Large-Scale Historical Textbases

2024

JDMDH, Journal of Data Mining & Digital Humanities

A Dataset for Named Entity Recognition and Entity Linking in Chinese Historical Newspapers

2024

LREC-COLING, Turin, Italy

Unlocking Transitional Chinese: Word Segmentation in Modern Historical Texts

2023

NLP4DH, Tokyo, Japan

Simulation d'erreurs d'OCR dans les systèmes de TAL pour le traitement de données anachroniques

2022

TALN (JEP-TALN-RECITAL), Avignon, France

Transferring Modern Named Entity Recognition to the Historical Domain: How to Take the Step?

2021

NLP4DH, Silchar, India

Creating Biographical Networks from Chinese and English Wikipedia

2021

Journal of Historical Network Research, Vol. 5, No. 1

Contextual Characters with Segmentation Representation for Named Entity Recognition in Chinese

2020

PACLIC 34, Hanoi, Vietnam