

Babak Khalilvandian

Data Scientist, Milan, Italy



khalilvandian



babak-khalilvandian



b.khalilvandian@gmail.com



+393517581189

EXPERIENCE

INTERN DATA SCIENTIST | KNIME

November 2024 – March 2025 | Konstanz, Germany

- **Generative AI & Agentic AI solutions:** Developed generative AI workflows, Agentic AI tools, and "Just KNIME It" materials.
- **Data app development & automation:** Built KNIME apps and dashboards with external API integrations.
- **ETL & geospatial pipelines:** Created KNIME workflows for ETL and geospatial data processing.
- **Data analysis & KPI extraction:** Extracted KPIs from LMS datasets.

FREELANCE/UNIVERSITY PROJECTS

ENTITY RECOGNITION AND LINKING ON GITHUB DATA | KHALILVANDIAN/ALLIGATOR

- **Built gold-standard dataset** — annotated GitHub company mentions with Wikidata QIDs (~500 entries) ensuring representativeness and evaluation reliability.
- **Developed and evaluated models** — applied Alligator neural entity linking framework with feature engineering and candidate reranking; explored fine-tuning using SemTab datasets.
- **Conducted error and feature analysis** — examined performance trade-offs, feature correlations, and data-domain mismatches; proposed improvements in type filtering and semantic feature design.

ABSTRACTIVE TEXT SUMMARIZATION BENCHMARKING (XSUM DATASET) |

KHALILVANDIAN/NLP-SUMMARIZATION

- **Implemented baseline and advanced models** — compared GRU-based Seq2Seq with Transformer architectures (T5, FLAN-T5, LLaMA, Qwen).
- **Applied fine-tuning strategies** — evaluated zero/few-shot prompting, full fine-tuning, and parameter-efficient tuning (LoRA, Prefix Tuning, IA3, Adapters).
- **Conducted performance analysis** — benchmarked with ROUGE and BERTScore; used Inseq for explainability to assess model behavior.

SOCIAL MEDIA OPINION MINING ON FORTUNE 1000 COMPANIES |

KHALILVANDIAN/FORTUNE_COMPANY_OPINION_MINING

- **Designed text-mining pipeline** — collected and enriched 590K Reddit/Twitter posts with Fortune 1000 sector and industry metadata.
- **Applied hybrid NLP methods** — performed topic modeling (LDA, anchored CorEx) and sentence-level clustering with transformer embeddings (all-mpnet-base-v2 + K-means).
- **Extracted interpretable insights** — identified consumer opinion facets (e.g., ethics, service, pricing) using coherence, Silhouette scores, and representative sentence analysis.

PUBLICATIONS

Novalija, Inna, et al. "From Code to Concept: A Semantic Approach to AI Innovation Discovery in Open Source Software Repositories." IEEE Access (2025).

EDUCATION

DATA SCIENCE MASTER'S |

UNIVERSITY OF MILANO-BICOCCA

2023 - 2025 | Milan, Italy

COMPUTER ENGINEERING BACHELOR'S | UNIVERSITY OF

TABRIZ

2017 - 2021

SKILLS

PROGRAMMING

Python (Primary) • KNIME • R • SQL • MQL

CORE DATA SCIENCE

ETL • Machine Learning • Data Visualization • Data Exploration • Data Cleaning • Statistical Analysis • Dashboard and Report Creation • KPI Analysis • KPI Analysis A/B Testing

SOFTWARE

Tableau • Docker • GitHub • Git • AWS • MongoDB • Apache Spark

SOFT SKILLS

Team Work • Analytical Skills • Critical Thinking

LANGUAGES

English - Level C2

Italian - Level B2

Persian - Native

Azerbaijani - Native

HOBBIES

Swimming and Classic Literature

Babak Khalilvandian

Data Scientist, Milano, Italia

ESPERIENZA

INTERN DATA SCIENTIST | KNIME

Novembre 2024 – Marzo 2025 | Costanza, Germania

- **Soluzioni di Generative AI e Agentic AI:** Sviluppo di workflow di Generative AI, strumenti di Agentic AI e materiali "Just KNIME It".
- **Sviluppo di app e automazione:** Creazione di applicazioni e dashboard KNIME con integrazione di API esterne.
- **Pipeline ETL e geospaziali:** Realizzazione di workflow KNIME per processi ETL e gestione di dati geospaziali.
- **Analisi dati ed estrazione KPI:** Estrazione di KPI da dataset LMS.

PROGETTI FREELANCE/UNIVERSITARI

ENTITY RECOGNITION AND LINKING SU DATI GITHUB

- **Creazione di un dataset gold-standard** — annotazione di menzioni aziendali su GitHub con QID di Wikidata (≈500 voci) garantendo rappresentatività e affidabilità della valutazione.
- **Sviluppo e valutazione di modelli** — applicazione del framework neurale Alligator per l'entity linking con feature engineering e reranking dei candidati; esplorata la fase di fine-tuning con dataset SemTab.
- **Analisi di errori e feature** — esame dei compromessi prestazionali, correlazioni tra feature e differenze di dominio; proposte migliorie in type filtering e progettazione di feature semantiche.

BENCHMARKING DEL RIASSUNTO ASTRATTIVO (DATASET XSUM)

- **Implementazione di modelli baseline e avanzati** — confronto tra Seq2Seq basato su GRU e architetture Transformer (T5, FLAN-T5, LLaMA, Qwen).
- **Applicazione di strategie di fine-tuning** — valutazione di prompting zero/few-shot, full fine-tuning e metodi di tuning a parametri efficienti (LoRA, Prefix Tuning, IA3, Adapters).
- **Analisi delle prestazioni** — benchmarking con metriche ROUGE e BERTScore; utilizzo di Inseq per analisi di explainability del comportamento del modello.

OPINION MINING SUI SOCIAL MEDIA PER LE FORTUNE 1000

- **Progettazione di una pipeline di text mining** — raccolta e arricchimento di 590K post da Reddit/Twitter con metadati su settore e industria delle Fortune 1000.
- **Applicazione di metodi NLP ibridi** — topic modeling (LDA, CorEx ancorato) e clustering a livello di frase con embedding transformer (all-mpnet-base-v2 + K-means).
- **Estrazione di insight interpretabili** — identificazione di opinioni dei consumatori (es. etica, servizio, prezzi) tramite coerenza, Silhouette score e analisi delle frasi rappresentative.

PUBBLICAZIONI

I. Novalija et al., "From Code to Concept: A Semantic Approach to AI Innovation Discovery in Open Source Software Repositories," under revision for publication in IEEE Access, 2025.

ISTRUZIONE

LAUREA MAGISTRALE IN DATA SCIENCE | UNIVERSITÀ

DEGLI STUDI DI MILANO-BICOCCA
2023 – 2025 | Milano, Italia

LAUREA TRIENNALE IN INGEGNERIA INFORMATICA

| UNIVERSITÀ DI TABRIZ
2017 – 2021

COMPETENZE

PROGRAMMAZIONE

Python (Principale) • KNIME • R • SQL • MQL

DATA SCIENCE DI BASE

ETL • Machine Learning • Data Visualization • Data Exploration • Data Cleaning • Analisi Statistica • Creazione di Dashboard e Report • Analisi KPI • A/B Testing

SOFTWARE

Tableau • Docker • GitHub • Git • AWS • MongoDB • Apache Spark

COMPETENZE TRASVERSALI

Lavoro di squadra • Capacità analitiche • Pensiero critico

LINGUE

Inglese – Livello C2

Italiano – Livello B2

Persiano – Madrelingua

Azero – Madrelingua