David Soares **Batista**

✉ dsbatista@gmail.com | ⌂ www.davidsbatista.net | davidsbatista | dsbatista

Summary

I'm an experienced machine learning engineer and software developer, with a strong background in Natural Language Processing. I'm skilled in applying Machine Learning, Deep Learning and other AI techniques to tackle diverse problems while also managing the associated software infrastructure ecosystem. My professional experiences span over academia, startup environments, freelancing projects, and established enterprises, where I've collaborated effectively with dedicated teams to deliver production-ready software solutions.

Professional Experience

Senior Data Scientist
Berlin, Germany

Veeva Systems GmbH
Mar 2023 - Present

- I conceived and developed within the Veeva Link product a model using PySpark MLlib to categorize life-science organisations using a 2 million-sample multilingual dataset and open data resources, reducing curator's workload by 10%. Additionally, I enhanced the researcher-to-activity linking model with pre-trained static embeddings in a 3.2 billion-sample annotated dataset, resulting in a 2% model performance improvement.
- Implemented Python development best practices (tests, linting, PEP 8, type hints) within a CI pipeline and mentored junior data scientists in applying the same practices.

Lead NLP Engineer
Berlin, Germany

TripActions/Navan (acquired Comtravo GmbH)
May 2022 - Feb 2023

- Joined through the acquisition of Comtravo by TripActions/Navan and the team was tasked with developing a chatbot's NLP/NLU component to automate customer support, incorporating previously developed components from the email channel into the new chatbot channel.
- Together with the team defined an intent and entities schema, set up the infrastructure for annotation based on the argilla.io annotation tool and onboarded annotators.
- Defined and implemented an entropy-based active-learning strategy to select samples for annotation.
- Fine-tuned, evaluated, performed threshold tuning and sequence tagging classifiers based on pre-trained Transformer architectures.

Lead NLP Engineer
Berlin, Germany

Comtravo GmbH
May 2021 - Feb 2022

- Led a team of 3 developers guiding them in technical decisions and supervised system changes, and also managed the 4 annotators.
- Managed the bi-weekly sprint planning and execution, coordinating development tasks based on system performance and feature requests.
- Defined quantifiable measures in collaboration with the Data Engineering team, resulting in performance reports and monitoring dashboards.
- Supervised the data annotation by ensuring annotation quality and consistency across the whole corpora.
- Continued to maintain an active role in software development: implementing new features, maintenance, refactoring and code reviews.

Senior NLP Engineer
Aug 2017 - Apr 2021

- Collaborated from it's inception in building an automation system to support travel agents in creating trip itineraries offers from incoming emails with business travel requests, working together with the team in multiple tasks.
- Trained, evaluated and deployed supervised models for text classification and sequence tagging using different architectures.
- Designed and constructed Knowledge Bases containing worldwide descriptions of airport and train stations based on open resources and in-house operational data.
- Conceived and implemented Entity-Linking approaches to map entities (e.g: airports, train stations, hotels) recognised in incoming travel requests into unique identifiers in Knowledge Bases.
- Collaborated together with the team in developing a ranking and selection algorithm of trip itineraries offers based on the Pareto Efficiency.
- Developed several modularised Python components with type-annotations, linting and test coverage of 95%.
- The automation system generating travel itineraries from customer emails reduced by 30% the workload of travel-agents.

Data Engineer
Berlin, Germany

HelloFresh SE
Jan 2016 - Jun 2017

- I was part of the initial Data Engineering team tasked with transitioning from crontab-based processes to an ETL platform.
- Developed the initial prototype for ETL management using Airflow, which evolved into the production data pipeline management system.
- Developed a classifier to identify customer feedback review mentioning issues with the ordered meal kits.

Researcher and Developer: REACTION - Computational Journalism
Lisbon, Portugal

INESC-ID Research & Development Institute
Jun 2011 - Apr 2014

- Explored and implemented methods for entity-relationship extraction, based on: hand-built patterns, supervised linear classifiers with linguistic features and semi-supervised methods based on seed relationships and large amounts of unannotated text.
- Developed an entity-linking approach associating personalities in news articles with Wikipedia using textual similarities and its graph structure.
- Built a graph based on topic models extracted from news articles and person co-occurrences, allowing to explore topics connecting persons.

Researcher and Developer: GREASE - Geographic Reasoning for Search Engines

Lisbon, Portugal

LaSIGE - Department of Informatics Research Unit

Sep 2008 - Oct 2010

- Developed a method to disambiguate toponyms based on textual context, information content and topological similarity measures.
- Built an alignment method between two geo-ontologies resulting in a single linked-data ontology, allowing the inclusion of features from both ontologies in SPARQL queries.
- Published scientific articles presenting the results of these tasks.

Software Developer

Lisbon, Portugal

Nokia-Siemens Networks

Oct 2007 - Jul 2008

- An 80% position while I took the remaining 20% to finish my M.Sc. curriculum component.
- Developed data collection modules for a GSM monitoring system using Java, CORBA Architecture and Oracle RDBMS.

Technical Skills

Programming	Python (`pytest, pylint, mypy`), Java, SQL, Unix-like Shell Scripting
Natural Language Processing Libraries	HuggingFace Transformers, spaCy, NLTK, gensim
Machine Learning Frameworks	scikit-learn, PyTorch, Keras
Information Retrieval	ElasticSearch, Kibana, Apache Lucene/Solr
Semantic Web/Linked Data	Apache Jena, SPARQL
Infrastructure	Docker, Flask, FastAPI, AWS: EC2, S3, Batch, EMR (PySpark)
Databases/Document stores	MySQL, PostgreSQL, MongoDB

Education

Doctor of Philosophy (Ph.D.) - “Large-Scale Relationship Extraction”

Lisbon, Portugal

Instituto Superior Técnico, University Lisbon

Feb. 2011 - Feb. 2015

- Researched different methods to perform semantic relationship extraction between named-entities and proposed a classifier based on the idea of nearest neighbour and leveraging min-hash and locality sensitive hashing for efficient similarity search.
- To obtain training data for the classifier I proposed a bootstrapping technique relying on distributional word representations which was awarded an *Honorable Mention for Best Short Paper* at **EMNLP’15**.

Master of Science (M.Sc.) - “Geographic Text Mining”

Lisbon, Portugal

Faculty of Sciences, University of Lisbon

Sep. 2007 - Jul. 2009

- Developed a geographic information extraction system based on Conditional Random Fields and a geographic ontology, to generate geographic summaries. The summary lists all the geographic entities found in a document mapped to unique concepts in the geographic ontology.
- The system was then applied to a crawl of the Portuguese Web (25GB raw text) using a Hadoop cluster, generating summaries for hundreds of thousands of documents.

Bachelor’s in Informatics Engineering

Lisbon, Portugal

Faculty of Sciences, University of Lisbon | Universität Karlsruhe, Germany

Sep. 2003 - Jul. 2007

- 2005-2006, Karlsruhe Universität (TH), Germany - 2 semesters as an exchange student
- 2004 November - 2005 July - IT support (part-time) - troubleshooting network infrastructure and preventative maintenance.
- 2003 November - 2004 March - Sys Admin (part-time) - administration and configuration of the IT department network and services.

Personal Projects

Politiquices.PT - Extraction of Support and Opposition Political Relationships

Sep 2020 - Apr 2021

- An award-winning project (2nd place, ”Arquivo.pt Awards 2021”) which extracts political interactions from Portuguese news headlines, it uses supervised models to detect interactions, linking the involved politicians to Wikidata entities creating an enriched RDF semantic graph. The graph allows for exploration by revealing nodes or graph clusters that depict the evolving political relationships of opposition or support between specific personalities, supported by information from news articles. I developed it during the 2020 pandemic while on short-time work, bringing to life an idea conceived during my Ph.D. - `https://www.politiquices.pt`

Bootstrapping Semantic Relationships with Distributional Semantics

Apr 2015 - Sep 2015

- `breds` is an open-source Python package stemming from my Ph.D. thesis, designed to extract named-entity relationships without labeled data. It utilizes an initial seed pairs of named-entities to learn extraction patterns, expanding and generalising relationships through distributional semantics while minimising semantic drift. - `https://pypi.org/project/breds`

Named-Entity Recognition Considering Partial Matching

May 2018 - Sep 2018

- `nervaluate` is an open-source Python package that evaluates named-entity recognition systems, including partial entity matching. It began as a proof-of-concept I worked on during my time at Comtravo and has since evolved into a mature Python package with contributions from various collaborators. - `https://pypi.org/project/nervaluate`