Alissa Ostapenko

aostapen@andrew.cmu.edu • https://ostapen.github.io

EDUCATION

M.Sc., Language Technologies **Expected August 2022**
Carnegie Mellon University (CMU) School of Computer Science • **GPA** 3.90/4.00 Pittsburgh, PA
Advisor: Dr. Yulia Tsvetkov

B.Sc. Computer Science & Mathematical Sciences, *Honors with High Distinction* **May 2020**
Worcester Polytechnic Institute (WPI)• **GPA** 3.93/4.00 Worcester, MA
Honors & Awards: *Salisbury Prize, Clare Boothe Luce Scholarship, Two Towers Prize, Presidential Scholarship*

SKILLS

Programming Languages: *Python 3, SQL, Java 8, MATLAB, C/C++, HTML, NodeJS, JavaScript*
Databases: *Oracle, PostgreSQL, MongoDB, Java DB/Apache Derby*
Frameworks & Toolkits: PyTorch, Pytorch-Lightning, *scikit-learn, pandas, numpy, plotly, spaCy, PIL, Flask, Docker*

RESEARCH EXPERIENCE

Graduate Research Assistant (Tsvetkov Lab) **Aug. 2020– Present**
Language Technologies Institute, Carnegie Mellon University Pittsburgh, PA

- Implemented a Transformer-based model to predict language switches in mixed English-Spanish dialogues using prior dialogue context and informative speaker prompts. Our model achieved performance boosts of up to 7% accuracy and 5 points F1 compared to speaker-agnostic baselines (PyTorch, PyTorch-Lightning)
- Incorporated and expanded an interpretability framework to highlight influential dialogue phrases, grounding our proposed model in linguistic and sociolinguistic literature
- We submitted our work to the ACL 2022 conference[1].

Undergraduate Research Assistant **January – May 2020**
Department of Computer Science, WPI Worcester, MA

- Through close collaboration with materials science PhD students and Dr. Danielle Cote (WPI), developed a flexible, easy-to-use Python API for predicting flowability of metal powders[5] (*Python 3, scikit-learn, pandas*)
- Implemented a variety of data preprocessing and feature extraction techniques, including correlation matrix-based feature selection and Principal Component Analysis, for optimizing model performance
- Expanded tool to allow for both regression (predicting a specific flowability value) and classification (predicting a range of flowability values), achieving perfect prediction accuracy for classification tasks
- Created well-documented Jupyter Notebooks for both the materials science team and for future student collaborators, allowing for reproducibility and easy code handoff

Data Science Intern **March – May 2019**
Vestigo Ventures, LLC & Dept. of Computer Science, WPI Cambridge, MA • Worcester, MA

- Developed FinDX, a machine-learning driven tool for classifying a company's business domain from its website, allowing Vestigo to identify new financial technology investment opportunities[6] (Python 3, NLTK, spaCy)
- Built novel web crawler and part-of-speech based parser, improving F1 classification score by 5% over Vestigo's previous baseline.
- Implemented a generalizable framework to extend classification to any business domain.

Undergraduate Researcher (Fifth Frederick Jelinek Memorial Summer Workshop) **June – August 2018**
The Johns Hopkins Whiting School of Engineering Baltimore, MD

- Implemented a supervised, hard-attention technique for multimodal machine translation, significantly improving alignments between words in the text and regions in the image[2,4] (PyTorch).
- Wrote a script to qualitatively evaluate image-text associations produced at model test time (Python 3).
- Presented research at a final presentation to industry sponsors including Google, Facebook, & Microsoft.

INTERNSHIP EXPERIENCE

Software Development Engineer Intern **May – August 2020**
Amazon Robotics, Manufacturing Division North Reading, MA

- Implemented foundational data processing and prediction pipelines for analyzing test performance of robotic drive units used in Amazon fulfillment centers

- Used statistical techniques to identify historical patterns in time series data from different drive unit components, to help flag suspicious behavior from new drive units before they are shipped for production
- Met with stakeholders to build dashboards for visualizing aggregated data from robot drive tests

Cognitive Software Engineering Intern **June – August 2019**
State Street Financial Services Boston, MA
- Core developer of orchestration service integrating Voice to Text and Amazon S3 storage services (Python3, Flask) between development teams in Hangzhou, China, Boston, MA, and Austin, TX.
- Designed and programmed database schema to track file and job metadata (PostgreSQL)
- Deployed service to State Street's internal cloud, by coordinating with the development team in Austin, TX
- Developed a sentiment analysis model for labeling positive and negative content within meeting transcriptions

Research Assistant Intern, Computer Vision **May – August 2017**
Newmetrix (previously Smartvid.io) Cambridge, MA
- Designed data collection experiments using Amazon Turk (AWS).
- Built multilayer perceptron classification models to identify and label target objects in an image (scikit-learn). Code was officially submitted into the company codebase.
- Utilized image processing techniques to analyze and label image data (Python 3, PIL) for training models and for evaluating their performance; presented results in biweekly Scrum meetings.

PUBLICATIONS

[1]***A Ostapenko**, S Wintner, M Fricke, Y Tsvetkov. Speaker Information Can Guide Models to Better Inductive Biases: A Case Study On Predicting Code-Switching. Submitted to ACL 2022*

[2]*L Specia, J Wang, SJ Lee, **A Ostapenko**, P Madhyastha: Read, spot and translate. Machine Translation (2021)*

[3]*S Arora*, **A Ostapenko***, V Viswanathan*, S Dalmia*, F Metze, S Watanabe, A W. Black: Rethinking End-to-End Evaluation of Decomposable Tasks: A Case Study on Spoken Language Understanding. Proc. Interspeech (2021)*

[4]*L Specia, L Barrault, O Caglayan, C Duarte, D Elliott, S Gella, N Holzenberger, C Lala, SJ Lee, J Libovický, P Madhyastha, F Metze, K Mulligan, **A Ostapenko**, S Palaskar, R Sanabria, J Wang, R Arora: Grounded Sequence to Sequence Transduction. IEEE J. Sel. Top. Signal Process. 14(3): 577-591 (2020)*

[5]*R Valente, **A Ostapenko**, B Sousa, J Grubbs, C Massar, D Cote, and R Neamtu: Classifying Powder Flowability for Cold Spray Additive Manufacturing Using Machine Learning: Proc. IEEE BigData (2020)*

[6]***A Ostapenko**, F Anderson, R. Neamtu: FinDX: A Versatile, Low Resource Approach to Financial Website Classification Proc. IEEE Big Data, December 2019.*

**: equal contribution*

ACTIVITIES & PROJECTS

Lead Developer (Mathematics & Computer Science Senior Capstone Project) **August – Dec. 2019**
Vestigo Ventures, LLC & Dept. of Computer Science ,WPI Cambridge, MA
- As the lead developer in a cross-disciplinary team of students, built the *Website Private Investigator (WPI),* a tool visualizing user navigation within company websites (Python3: *Plotly*, *NetworkX*, *pandas*)
- Designed and developed core architecture; extended tool with interactive analysis features
- Integrated the *WPI* into the company's data analysis portal, in collaboration with lead software engineers
- Researched, analyzed, and integrated clustering techniques to improve user flow visualizations (*scikit-learn*)

Student Coordinator (Iceland Project Center Initiative) **Spring & Fall 2018**
Department of Integrative & Global Studies, WPI WPI & Reykjavik, Iceland
- In a team of four, met with directors of nonprofits, government organizations, and museums to establish connections for future social science project partnerships between WPI students and Icelandic organizations
- Recommended accommodations, budgeting, tourism, and travel logistics for future WPI students completing projects in Reykjavik.

Teaching Assistant (Object-Oriented Programming; Linear Algebra II) **October 2018 – March 2019**
Honor Societies: Pi Mu Epsilon, Upsilon Pi Epsilon