Krishnapriya Vishnubhotla

Room 276, D.L. Pratt Building, 6 King's College Road, Toronto, Ontario – M5S 3H5

✉ vkpriya@cs.toronto.edu • https://priya22.github.io/

I am a second-year PhD student in the Computational Linguistics group at the University of Toronto. My main research area is **Natural Language Processing**. I am interested in developing techniques that can help machines interpret, understand and participate in human communication. I am currently supervised by **Prof. Graeme Hirst**, and co-supervised by **Prof. Frank Rudzicz**, of the Department of Computer Science, University of Toronto.

Education

- **University of Toronto** — **Toronto**
  PhD in Computer Science, Computational Linguistics Group, GPA: 3.97/4.0 — *2019–Present*
- **University of Toronto** — **Toronto**
  Master of Science in Computer Science, Thesis option, GPA: 4.0/4.0 — *2017–2019*
- **National Institute of Technology Karnataka-Surathkal** — **Mangalore, India**
  B.Tech Computer Science and Engineering , CGPA 8.93/10 — *2013–2017*
- **K.V.I.I.Sc** — **Bangalore, India**
  Primary, Secondary and High School (Grades 1 to 12), 97% AISSCE — *2001–2013*

Work Experience

- **Georgian Partners** — **Toronto**
  Research Intern — *May 2020 – Present*
  As a research intern at Georgian, I work on developing unsupervised clustering models of text embeddings for internal company applications.
- **Samsung AI Research Center** — **Toronto**
  Research Intern — *May 2019 – September 2019*
  As a research intern at Samsung AI, I worked on multi-modal representation learning and semi-supervised methods of text and video alignment. I was a part of the winning submission for the Samsung Retail Robot Challenge, for which we built an interactive clip retrieval system for customer support videos.
- **University of Toronto** — **Toronto**
  Teaching Assistant — *September 2017–Present*
  Teaching Assistant for the following courses. My duties include preparing tutorials, quizzes and taking lab hours.
  - CSC401/2511: Natural Language Computing.
  - CSC108: Introduction To Computer Programming.
  - CSC309: Programming on the Web.

Notable Projects

- **Annotation Tool for Character Dialogue in Literary Texts (Sept 2019-Current)**

  Developed a web-based tool for annotating quotations and associated mentions in a text. The tool allows users to select and annotate quotations for speaker, addressee(s), referring expression, and other metadata. This integrates with the *GutenTag* tool to automatically extract character names and quotations in text.

- **Algorithms for Private Data Analysis: Course Project (Fall 2019):** *'Author Profiling and Privacy for Text'*

  Investigated the effeciveness of an RNN-based autoencoder in conjunction with a gradient reversal layer for the task of obfuscating stylistic markers of author identity. We tested the method on the two authorship attribution corpora, and showed that the model succesfully reduced the accuracy of an SVM-based style classifier by 30% on average. We demonstrated the difficulty of training such networks on small and unbalanced datasets.

- **Masters Thesis Project (2018-2019):** *'A Stylometric Investigation of Character Voices in Literary Fiction'*

  We investigate the stylistic characteristics of character dialogue in a corpus of modern-era plays. We demonstrate the effectiveness of Bayesian lexical models combined with word embeddings on this task, and uncover some interesting commonalities in character voices across authors and texts. We further use our stylistic features to build a semi-supervised quote attribution algorithm that achieves performance comparable to the state-of-the-art, while requiring lesser manual annotation.
- **Learning Dicrete Latent Structure: Course Project (Winter 2018):** *'GANs for Text Generation using word2vec'*

  Developed a text generation model using Generative Adversarial Networks (GANs), with word2vec embeddings as the input and output of the system. This bypasses the problem of differentiating through a discrete space, i.e, words. Both the generator and the dicriminator were implemented as deep convolutional layers. A conditional variant of the model was also implemented.
- **CLEF 2018 eHealth Challenge Task 1 (Winter 2018):** *'Multi-lingual ICD-10 Coding using an Ensemble of Recurrent and Convolutional Neural Networks.'*

  We assign ICD-10 codes to cause-of-death phrases in multiple languages by creating rich and relevant word embedding models. We train 100-dimensional word embeddings on the training data provided, combined with language-specific Wikipedia corpora. We then use n-gram matching of the raw text to the provided ICD dictionary followed by an ensemble model which includes predictions from a CNN classifier and a GRU encoder-decoder model.
- **Masters Project (Fall 2017):** *'Extracting opinions from Online Reviews'*

  Worked on opinion mining and sentiment analysis of reviews in the healthcare domain. We looked at two datasets: the first deals with performance reviews of medical residents written by their supervisors, and the second with reviews of doctors written by thier patients on the RateMDs platform. I looked into automated, unsupervised methods of detecting opinion and target terms, as well the associated sentiment, from these reviews. Some of the methods explored were topic models and phrase mining techniques.

Publications

- Hammond, A., Hirst, G. and Vishnubhotla, K., 2020. The Words Themselves: A Content-based Approach to Quote Attribution. *Digital Humanities 2020*
- Vishnubhotla, K., Hammond, A. and Hirst, G., 2019, June. Are Fictional Voices Distinguishable? Classifying Character Voices in Modern Drama. In *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature* (pp. 29-34).
- Budhkar, A., Vishnubhotla, K., Hossain, S. and Rudzicz, F., 2019, August. Generative Adversarial Networks for Text Using Word2vec Intermediaries. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)* (pp. 15-26).
- Jeblee, S., Budhkar, A., Milic, S., Pinto, J., Pou-Prom, C., Vishnubhotla, K., Hirst, G. and Rudzicz, F., 2018. TorontoCL at the CLEF 2018 eHealth Challenge Task. In *CLEF 2018 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS, September 2018.*

Relevant Courses

○ (Advanced) Computational Linguistics ○ Natural Language Computing ○ Introduction to Machine Learning ○ Learning Discrete Latent Structure ○ Algorithms for Private Data Analysis ○ Topics in Computational Social Science ○ Computational Models of Semantic Change

Technical skills

- **Machine Learning Frameworks:** PyTorch, TensorFlow, FastAI, HuggingFace
- **Programming Languages:** Proficient in: C, C++, Python, Javascript, Ruby
  Basic ability with: Java.
- **Industry Software Skills:** SQL, LaTeX, MATLAB, Ruby on Rails