LIFE TRAJECTORIES AS SYMBOLIC LANGUAGE

Exploring Human Behaviour with Language Models

PhD Thesis
Germans Savcisens

GERMANS SAVCISENS

LIFE TRAJECTORIES AS SYMBOLIC LANGUAGE

TECHNICAL UNIVERSITY OF DENMARK

COVER: Germans Savčišens, December 2023

PUBLISHED BY: Technical University of Denmark,
Department of Applied Mathematics and Computer Science
Richard Petersens Plads
Building 324
2800 Kgs. Lyngby
Denmark
`www.compute.dtu.dk`

First printing, December 2023

Statement of Approval

This thesis is submitted in partial fulfillment of the requirements for obtaining the Doctor of Philosophy (Ph.D.) degree at the Technical University of Denmark (DTU).

The Ph.D. project was conducted at the Section for Cognitive Systems at the DTU Department of Applied Mathematics and Computer Science between 1 September 2020 and 30 November 2023. It was conducted in collaboration with the Copenhagen Center for Social Data Science at the University of Copenhagen and the Data Science Lab at Statistics Denmark. The project was supervised by Professor Sune Lehmann Jørgensen (primary supervisor) and Professor Lars Kai Hansen.

Additionally, a part of the Ph.D. project was conducted at the Network Science Institute, Northeastern University, under the supervision of Professor Tina Elliasi-Rad, between 1 September 2022 and 28 February 2023.

The Ph.D. project received funding from Villum Fonden as a part of the *Nation-scale Social Networks* project. Further, the Ph.D. project was supplemented by travel grants from Stibo Fonden, Otto Mønsted Fonden, and William Demant Fonden.

Kongens Lyngby, December 2023
Germans Savčišens

This is for my mum and childhood friends, Lera and Zhenya.

Acknowledgments

I thank my family and close friends for always being around during this journey. I extend my deepest gratitude to Sune Lehmann for making this project a reality and providing relentless energy and exceptional mentorship over the years. I am equally grateful to the members of the Cognitive Science section, especially Lars Kai Hansen, whose profound knowledge and insights have been invaluable. A special thank you goes to my work buddies, Agata Wlaszczyk and Nikolaos Nakis, for keeping me company, especially during the earlier days of the pandemic – I still think about our virtual coffee meetings. I also express my heartfelt appreciation to the Social Complexity Lab for their unwavering support, guidance, and the positive energy that fueled my research endeavors. Gratitude is also due to Laust Mortensen and the dedicated team at Denmark Statistics for their assistance and swift support. I am indebted to Tina Eliassi-Rad and the Network Science Institute at Northeastern University for hosting me during my external research stay. I am profoundly thankful to my co-authors, including Lau Lilleholt, Anna Rogers, and Ingo Zettler, for their substantial contributions to the `life2vec` paper. I also wish to thank Thomas Jones for not only hosting me in Boston but also for enlightening discussions about transformers, proteins, and T-cells. My appreciation extends to Roberta Sinatra and the team behind the Algorithmic Fairness, Accountability, and Ethics course for offering me a valuable teaching opportunity. A special acknowledgment to the Copenhagen Center for Social Data Science at the University of Copenhagen for their warm welcome and support. Lastly, a big thank you to Aske Mottelson for his excellent translation of my abstract into Danish.

Contents

PROVENANCE

UNFOLDING

INSTRUMENT

DIVINATIONS

Acronyms

AUC Area Under the Curve. 129

AUL Area Under the Lift. 130, 133, 136, 137

BERT Bidirectional Encoder Representations from Transformers. 15, 58, 62, 79, 80, 91, 92, 94, 96, 97, 99, 103, 106, 108, 111, 124

BoW Bag-of-Words. 38, 40, 41, 136, 137

C-MCC Corrected Matthews Correlation Coefficient. 130, 131, 136, 137, 139, 140

CAV Concept Activation Vector. 148, 150

CBOW Continuous Bag-of-Words. 46

CDW-CE Class Distance Weighted Cross-entropy. 154, 156

CNN Convolutional Neural Networks. 63

CORAL Consistent Rank Logits. 155

CORN Conditional Ordinal Regression for Neural networks. 155

CoV Coefficient of Variation. 155, 156

CPR Civil Registration Number. 35, 68

*DB*07 Danish Industrial Classification of All Economic Activities 2007. 17, 70, 76, 117

DST Statistics Denmark. 35, 76

ESA-2010 European System of Accounts. 70, 76

FPR False Positive Rate. 130, 131

GLoVe Global Vectors for Word Representation. 46

GLU Gated Linear Unit. 52

GNN Graph Neural Network. 56, 57, 63

GRU Gated Recurrent Unit. 47, 48, 84, 137, 161–163

ICD-10 International statistical classification of diseases and related health problems, 10th revision. 17, 72, 73, 76, 115–117, 151

IDF Inverce Document Frequency. 41

IELTS International English Language Testing System. 94

ISCO-08 International Standard Classification of Occupations. 17, 71, 76, 115, 117, 151

k-NN k-Nearest Neighbors. 38

LLM Large Language Models. 23, 24, 32, 33, 43, 52–54, 57, 60, 62, 99, 134

LSTM Long Short-Term Memory. 47, 107

MCC Matthews Correlation Coefficient. 130, 131

MIDI Musical Instrument Digital Interface. 59, 60, 75

MLM Masked Language Model. 15, 58, 77, 92–99, 109, 113, 134, 135

MSE Mean Squared Error. 155

NACE Nomenclature des Activités Économiques dans la Communauté Européenne. 70

NLP Natural Language Processing. 32, 33, 36, 40, 52, 59, 62, 75, 77, 93–95, 112, 134

NSP Next Sentence Prediction. 58, 97

ORF Orthogonal Random Features. 105

PaCMAP Pairwise Controlled Manifold Approximation. 114, 116, 144

PCA Principal Component Analysis. 41, 105

PID Personal ID. 68, 69

POSAP Det danske personligheds og sociale adfærdspanel. 84, 87

PU Positive-Unlabeled. 129–133, 143, 154

PWFF Position-wise fully connected (feed forward) layer. 106–108, 111

QWK Cohen's Quadratic Kappa Score. 160, 161

ReLU Rectified Linear Unit. 51, 106

RNN Recurrent Neural Network. 47, 48, 52, 53, 55, 56, 58, 63, 77, 107

SNA-2008 System of National Accounts 2008. 70

SOP Sequence Order Prediction. 15, 92, 96–99, 109, 110, 113, 134, 135

TCAV Testings with Concept Activation Vectors. 17, 148–152, 163

TF Term Frequencies. 17, 38–41

TF-IDF Term Frequency - Inverse Document Frequency. 41

TOEFL Test of English as a Foreign Language. 94

UMAP Uniform Manifold Approximation and Projection. 41, 115

UNSC United Nations Statistical Committee. 70

WHO World Health Organization. 72, 73

List of Figures

List of Tables

Abstract

Deep learning has significantly advanced research within natural language processing in recent years. Beyond language, novel transformer-based architectures have shown promise as tools to model various multivariate sequences. These include weather patterns, musical compositions, and protein structures. Similarly, human lives represent another form of multivariate sequences comprising various events: People are born, attend kindergarten, visit doctors, relocate to new cities, and more.

By drawing parallels between human lives and written language, we propose a novel methodology to study individual life trajectories. We use the Danish National Registry to create an artificial symbolic language. It transforms socioeconomic and health events into a structured, sentence-like format, akin to the words and sentences in a language.

This representation approach lays the groundwork for our primary contribution: developing the `life2vec` model, a transformer-based model designed for analyzing life trajectories. In the thesis, we demonstrate that the `life2vec` model captures complex relationships between various life events and uses this knowledge to provide insights into early mortality and personality. A key strength of `life2vec` lies in interpretability, as we can use it to explore the influence of socioeconomic and health factors on individual life paths.

The findings from this thesis underscore the potential of transformer models in understanding and predicting human behavior and experiences.

Resumé

Deep learning har i de seneste år fremmet forskningen i naturlig sprogbehandling betydeligt. Udover sprog, har nye transformer-baserede arkitekturer vist sig lovende som værktøjer til at modellere forskelligartede multivariate sekvenser. Disse inkluderer vejrmønstre, musikalske kompositioner og proteinstrukturer. Menneskeliv repræsenterer på samme vis en form for multivariate sekvenser, der omfatter forskellige begivenheder: Folk bliver født, går i børnehave, går til læge, flytter til nye byer og mere.

Ved at drage paralleller mellem menneskeliv og det skrevne sprog foreslår vi en ny metodik til studiet af individuelle livsforløb. Vi benytter Folkeregisteret til at skabe et kunstigt symbolsk sprog. Det transformerer socioøkonomiske og sundhedsmæssige begivenheder til et struktureret, sætningslignende format, sammenligneligt med ordene og sætningerne i et sprog.

Denne repræsentationstilgang danner grundlaget for vores primære bidrag: udviklingen af life2vec-modellen; en transformer-baseret model designet til at analysere livsforløb.

I denne afhandling demonstrerer jeg at life2vec-modellen fanger komplekse sammenhænge mellem forskelligartede livsbegivenheder og benytter denne viden til at give indsigter i tidlig dødelighed og personlighed. En vigtig egenskab ved life2vec ligger i fortolkningsgraden, da vi kan bruge den til at udforske indflydelsen af socioøkonomiske og sundhedsmæssige faktorer på individuelle livsforløb.

Resultaterne fra afhandlingen understreger transformer-modellernes potentiale til at forstå og forudsige menneskelig adfærd og oplevelser.

Preface

The way I wrote the thesis is partly inspired by the stats book by Field and Hole[1]. It was mandatory reading for one of the courses I studied during my bachelor studies in Medialogy at Aalborg University. That is how I got into statistics and machine learning in the first place. How could you not after reading about the drug-induced cats and hats forgotten in a freezer?

Scope of the thesis. The material of the thesis covers *one* publication developed during the Ph.D. project, namely

> **Savcisens, Germans**, Tina Eliassi-Rad, Lars Kai Hansen, Laust Hvas Mortensen, Lau Lilleholt, Anna Rogers, Ingo Zettler, and Sune Lehmann. "Using Sequences of Life-events to Predict Human Lives." Accepted at *Nature Computational Science* (2023). *DOI: 10.1038/s43588-023-00573-5*

The thesis **does not cover** other publication developed during the Ph.D project

> Fernández, Elena Fernández, and **Germans Savcisens**. "A Sustainable West? Analyzing Clusters of Public Opinion in Sustainability Western Discourses in a Collection of Multilingual Newspapers (1999-2018)." *Digital Humanities in the Nordic and Baltic Countries Publications* 5, no. 1 (2023): 165-187.

> Denove, Emmanuelle, Elisa Michelet, **Germans Savcisens**, and Elena Fernández Fernández. "An Industrial West? A Quantitative Analysis of Newspapers Discourses about Technology over Ninety Years (1830-1940)." Preprint at `http://doi.org/10.5281/zenodo.8255722` (2023).

The peer-reviewed and accepted version of the "*Using Sequences of Life-events to Predict Human Lives*" paper is available in the Appendix materials, along with the peer-reviewed and accepted version of the "*A Sustainable West? Analyzing Clusters of Public Opinion in Sustainability Western Discourses in a Collection of Multilingual Newspapers (1999-2018)*" paper.

Purpose of the thesis. I wrote this thesis as a series of hypothetical lecture notes on using Large Language Models (LLM) to explore

[1] A. Field and G. Hole. *How to Design and Report Experiments*. Sage, 2002

socioeconomic and health life-trajectories.

I wrote it so that any person with a basic understanding of statistics and basic knowledge of machine learning (but who might still be uncomfortable within the areas of natural language processing, representation learning, and deep learning) could still understand the motivation and the theory behind decisions and results of the `life2vec` model.

Secondly, the fact that LLMs become a more prominent part of our lives also contributed to how I present the material. LLMs seem to be all-powerful models and, in the eyes of many, are surrounded by a certain *mystic* added by the over-promising tech industry – but LLMs do have limitations.

I take this opportunity not only to present how we can use LLMs to look at human lives from a new viewpoint but also to demonstrate what LLM are essentially capable (and not capable of).

Main question. The question that I explore in my thesis is: *How do we adapt the neural representation methods to model human behavior and experiences?* To answer this, my research focuses on two key aspects:

1. use of the **National Registry Data** to study socioeconomic and health trajectories,
2. using **Transformer**-based models to capture and interpret phenomena in these life trajectories.

As I demonstrate in the thesis, life trajectories are conceptually similar to sequential signals such as text, music, or protein structure. As one can understand: "Sage got a full position as an Insurance Agent after many years of work as a Realtor," so one can transcribe it using simple symbols to communicate similar information, and so one can teach a computer to *understand* it.

Structure of the thesis In the Provenance chapter, I provide the motivation behind the use of the Danish Registry data for the exploration of human behavior. I also provide the motivation and theory behind the use of the transformer-based models.

In the Unfolding chapter, I describe what kind of data we use to analyze socio-economic and health life trajectories. I also provide details on the artificial symbolic language that transcribes tabular data into the life-sequence.

In the Instrument chapter, I provide a detailed description of the `life2vec` model and the results of the pretraining of `life2vec`. I show how we can use it to study relationships between various life-events.

In the Divinations, I show how we use `life2vec` for specific tasks such as mortality prediction and personality nuance prediction tasks.

This chapter also describes the method for the `life2vec` explainability.

The final chapter, Coda, contains a short work summary.

Use of the Generative AI. During the development of the thesis, I used the following platforms and models for image generation, information search, text formatting, and proofreading:

- `MidJourney`,
- `Adobe Firefly`,
- Microsoft `Copilot` with Bing Chat,
- `GPT-4` (including "Research by Vector" plugin),
- `DALLE-3`,
- `GrammarlyGO`.

Writing. I use "we" and "I" interchangeably to improve the flow. You will also notice that I use commas as a *thousands separator*, *e.g.*, $1,000$ or $1,000,000$. I utilize italicized symbols, such as s, to specify scalars, and bold symbols, like $\boldsymbol{Z}$ and $\boldsymbol{z}$, to denote matrices or vectors. I use this symbol $\cdot$ for the dot product, this symbol $\times$ exclusively for the scalar multiplication, and this symbol $\odot$ for the element-wise multiplication.

And finally, I know it gets a little bit silly in some places, mainly due to my love of arbitrary quotes – but I prefer to keep it that way. Enjoy!