Text Mining with DSL

Sukanya Nath

Ahmad Alhineidi

Image generated by DALLE openai

Data Science Lab (DSL)

- DSL is an interfaculty core facility of the University of Bern.
- We provide university-wide support and training for researchers and research groups in data science, machine learning, artificial intelligence and research IT related matters.
- Our Services range from rapid advice by e-mail and code or algorithm reviews to long-term collaborations on infrastructure and research projects with co-analysis and co-authoring.
- We Consolidate several research support activities (Digital Humanities, Science IT Support, Computer Vision, Text Mining,…)
- UniBe analogue to ETHZ SIS, UZH S3IT, UniBas SciCORE, EPFL, SDSC, …
- Mandate from 2023-01-01 to 2026-12-31 (prolongation upon positive evaluation)

Interdisciplinary Team

(domain experts, technical experts and data scientists)

Sigve Haug
Coordinator / All Rounder

Guillaume Witz
Computer Vision Support / Training

NLP Support / ChatGPT & Co

Alexander Kashev
Research IT Support

Mykhailo Vladymyrov
Machine Learning Support

Ana Stojiljkovic
Computer Vision Support

Ninoska Friedli
Administrative Coordinator

Stephen Hart
Digital Humanities/ History

Peter Dängeli
Digital Humanities Support

Marco Indermühle
HPC System Administrator

Ahmad Alhineidi
NLP and HPC Support

Sebastian Flick
Frontend developer. Digital Hum. support

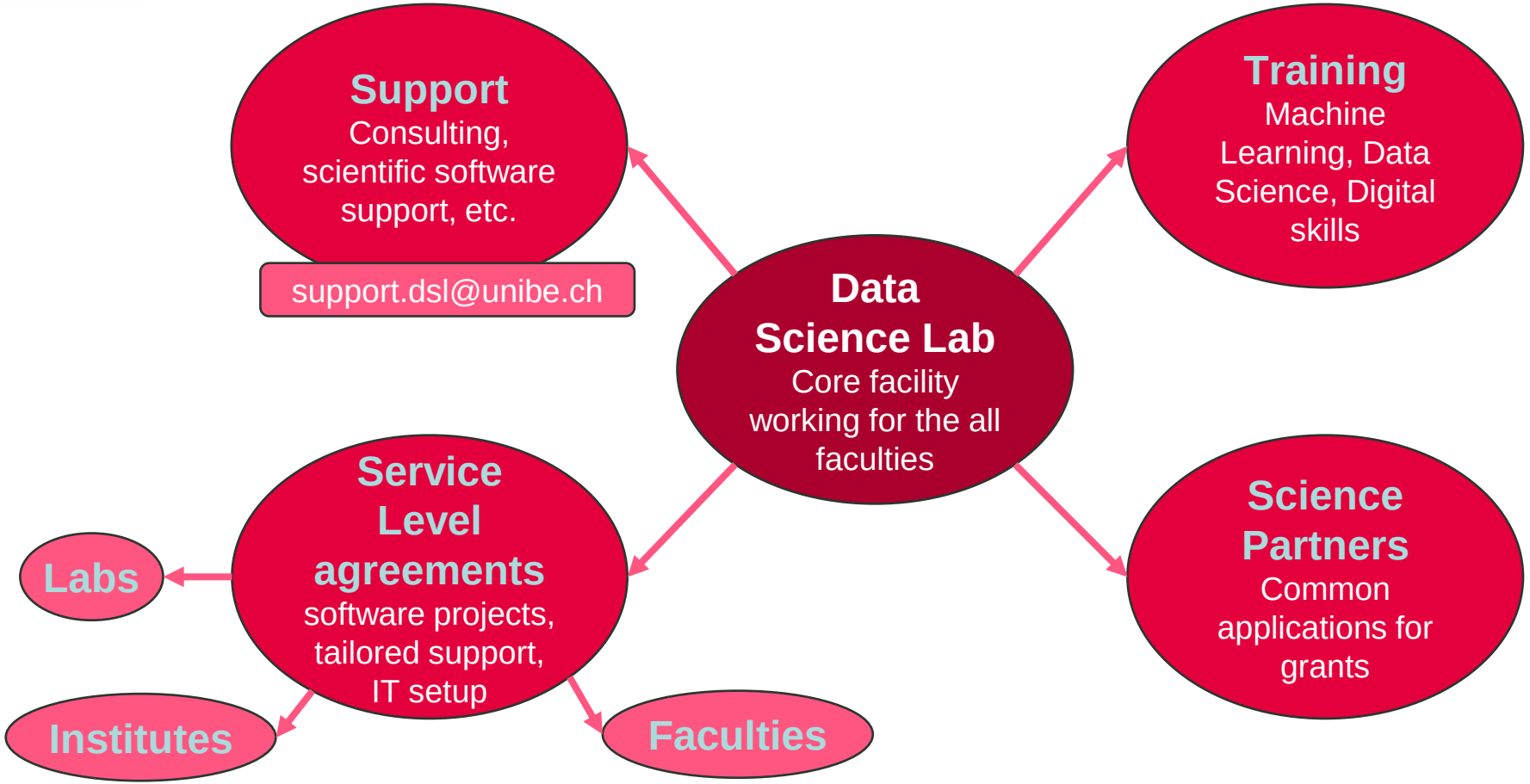
Sebastian Borkowski
Digital Humanities Support

Sukanya Nath
NLP Support

Michael Horn
Computer Vision Support

Assistants

What is DSL?


Support
Consulting, scientific software support, etc.
support.dsl@unibe.ch
Training
Machine Learning, Data Science, Digital skills
Data Science Lab
Core facility working for the all faculties
Service Level agreements
software projects, tailored support, IT setup
Science Partners
Common applications for grants
Labs
Institutes
Faculties

DSL Walk-in Tuesdays

Do you have trouble running a data processing script?

Do you want to discuss how to use Machine Learning in your project?

Do you have any other question on Data Science or Research Software and IT?

Just come by the Data Science Lab and we will try to help you on the spot!

Data Science Lab Walk-In

Where: Main building (HS4), rooms 311 - 312

When: Tuesdays 13.00 - 15.00

DSL Trainings

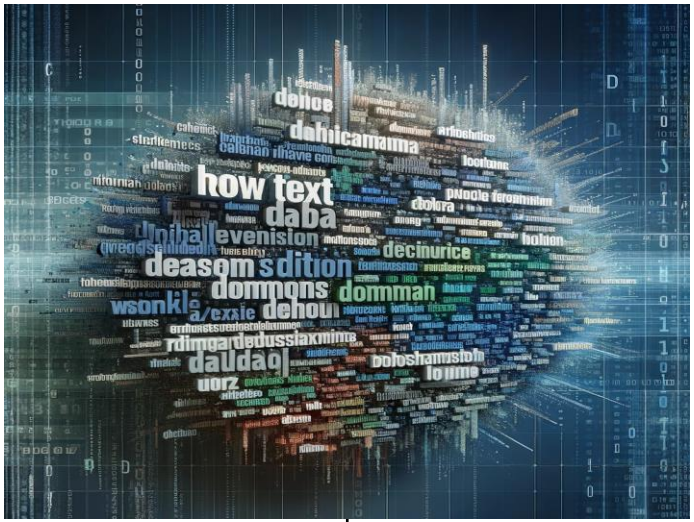
Winter schools

- Bern Winter School on Deep Learning
- Bern Winter School on Reinforcement Learning
- Bern Winter School on Natural Language Processing

Certificate of Advanced Studies

- CAS Natural Language Processing
- CAS Applied Data Science
- CAS Advanced Machine Learning
- CAS AI for Creative Practices

Text Mining Overview

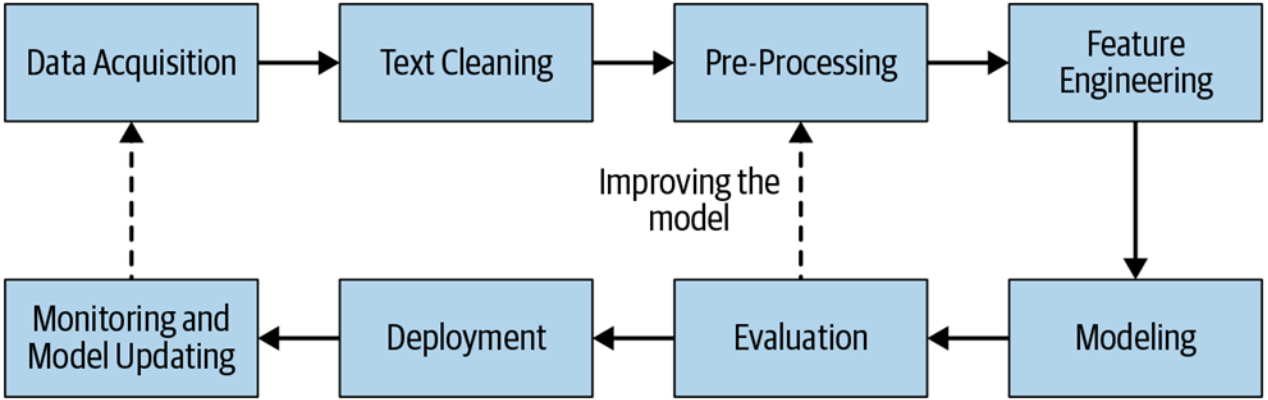


Knowledge!!

Image generated by DALLE openai

Text Mining Overview

Generic Text Mining Pipeline



Source: Vajjala et al. 2020

Text Mining Overview

Text Mining problems

- Classification
- Summarization
- Information Extraction
- Named Entity Recognition
- Question Answering
- Topic Modelling
- Key Phrase Extraction
- Relation Extraction
- Syntactic/ Linguistic Feature Extraction
- Data Annotation/ Augmentation/ Anonymization

Text Mining Classification

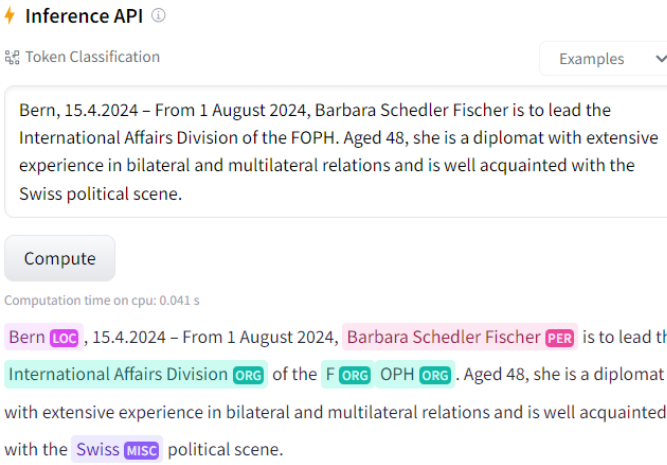
- Binary or Multi-class classification
- Supervised ML

Example Usage:

- Sentiment analysis: Do people have a favourable opinion of the government?
- Hate speech detection: How does hate speech differ from free speech?
- Customized task: Is this text written by a man, woman or a bot?
- Spam detection
- Language identification

Text Mining Classification

- Token Classification (NER, POS)
- For Information Extraction
- Linguistics features and analysis



Text Mining Summarization

- Sequence-to-sequence models.
- Helpful for further steps such as text classification.
- Extractive or Abstractive.

Example usage:

- You want to know the change in stance of the UN with respect to a certain topic over time.
- You have access to the speeches made by UN secretary generals but these speeches are large, numerous and can contain topics irrelevant for your research.
- Summarization can help to condense the content of the speeches. It may also be possible to guide it in a certain direction.

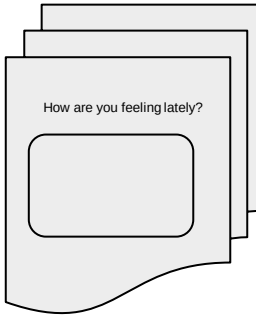
Image generated by DALLE openai

Text Mining Topic Modeling

- Unsupervised learning
- Based on clustering algorithms
- Extract knowledge from massive amount of documents

Example usage:

- You have performed an open ended survey of people suffering from anxiety.
- You want to know what topics were commonly mentioned.
- Topic Modelling can help to identify the topics and can also be guided towards a preferred direction.

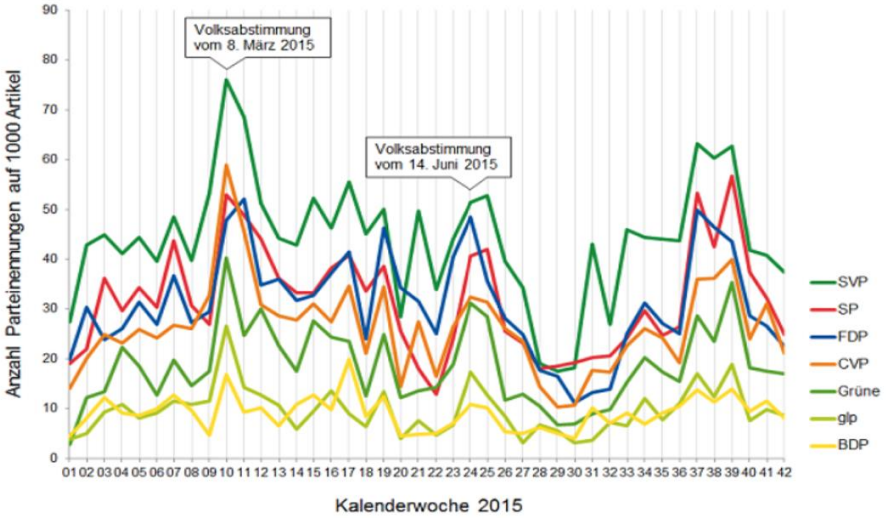


Data Annotation/ Augmentation/ Anonymization

- Data Annotation/ Labelling is the process of identifying and labelling the context of the text.
- Data Augmentation is the process of artificially generating newer data points from existing ones.
- Data Anonymization is the process of removing personally identifiable information from the data set.

Text Mining Output Example

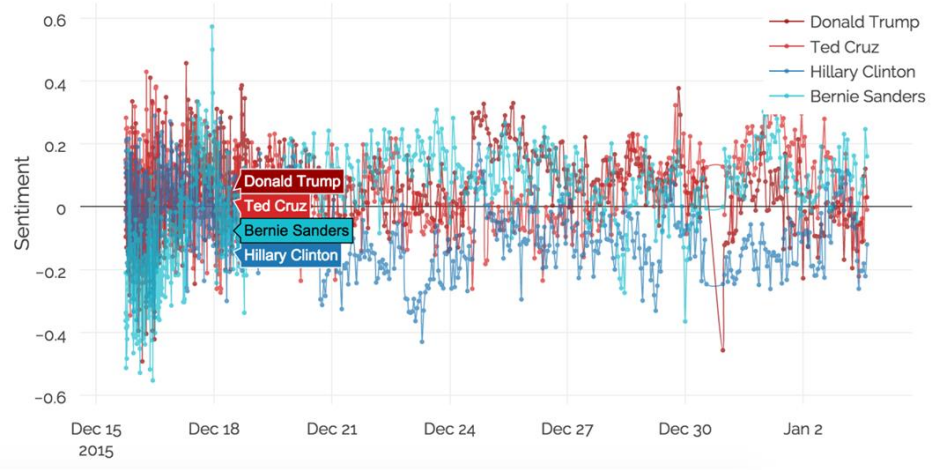
- Number of mentions of Swiss Political parties
- How to achieve such text mining output?



Source: Eurospider Newsletter

Text Mining Output Example

- Twitter data on sentiment of presidential candidates in 2015
- How to achieve such text mining output?



Source: Election Analysis and Data Science. link

Closing Thoughts

- Text Mining methods have improved massively over the past few years especially with the advent of generative AI.

- However
  - Models / Data may be prone to bias/ stereotypes.
  - Access to data must respect privacy laws.
  - Explaining and visualising the decisions by models is important but can be difficult.

THANK YOU!

QUESTIONS?