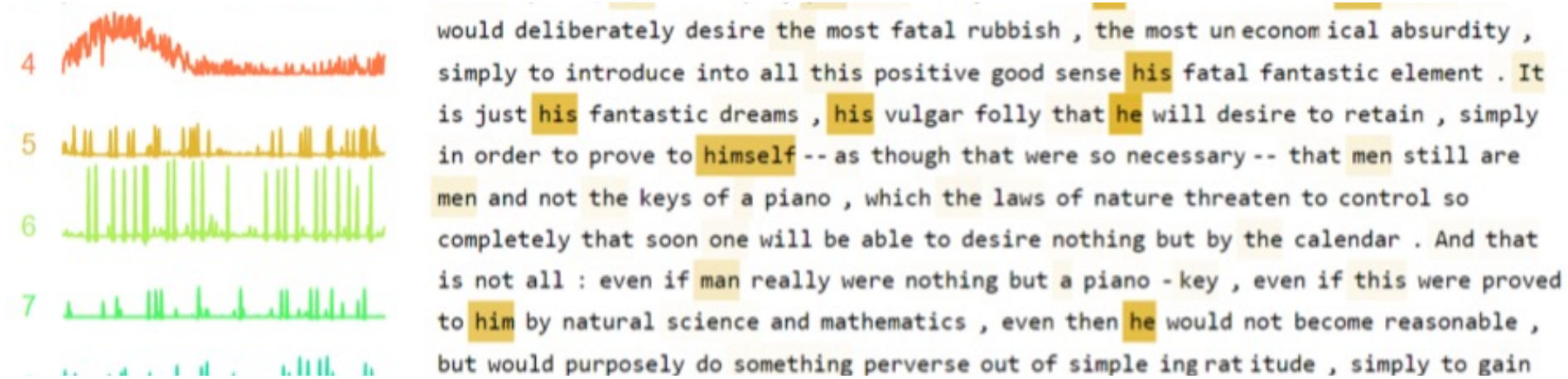


VU Visualisierung 2 (186.833)

Exploratory Error Analysis of Named Entity Recognition Models and Datasets

Alexander Seifert, 0326507

Institute of Visual Computing & Human-Centered Technology, TU Wien, Austria

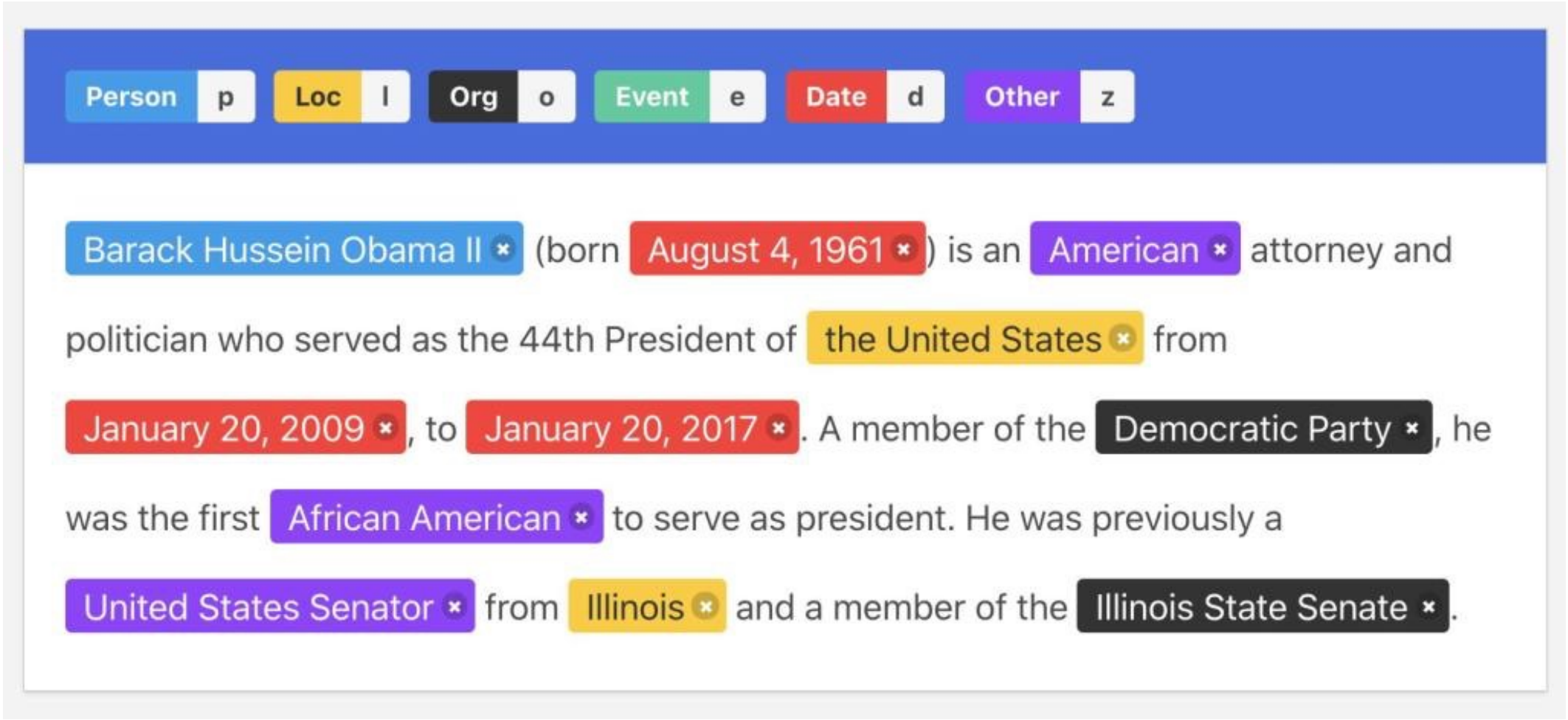
Motivation

- State of Deep Learning & Natural Language Processing in 2022:
  - Deep Learning **models** are still **poorly understood** *(black box)* [1]
  - Growing recognition of **data quality importance** *(data-centric AI)* [2]
  - Recent studies have shown there's a considerable number of **labeling errors** in standard benchmark datasets (3.4% avg over 10 datasets) [3]
- **Error Analysis** is a super-important but often overlooked part of the data science project lifecycle, both **for models & datasets**

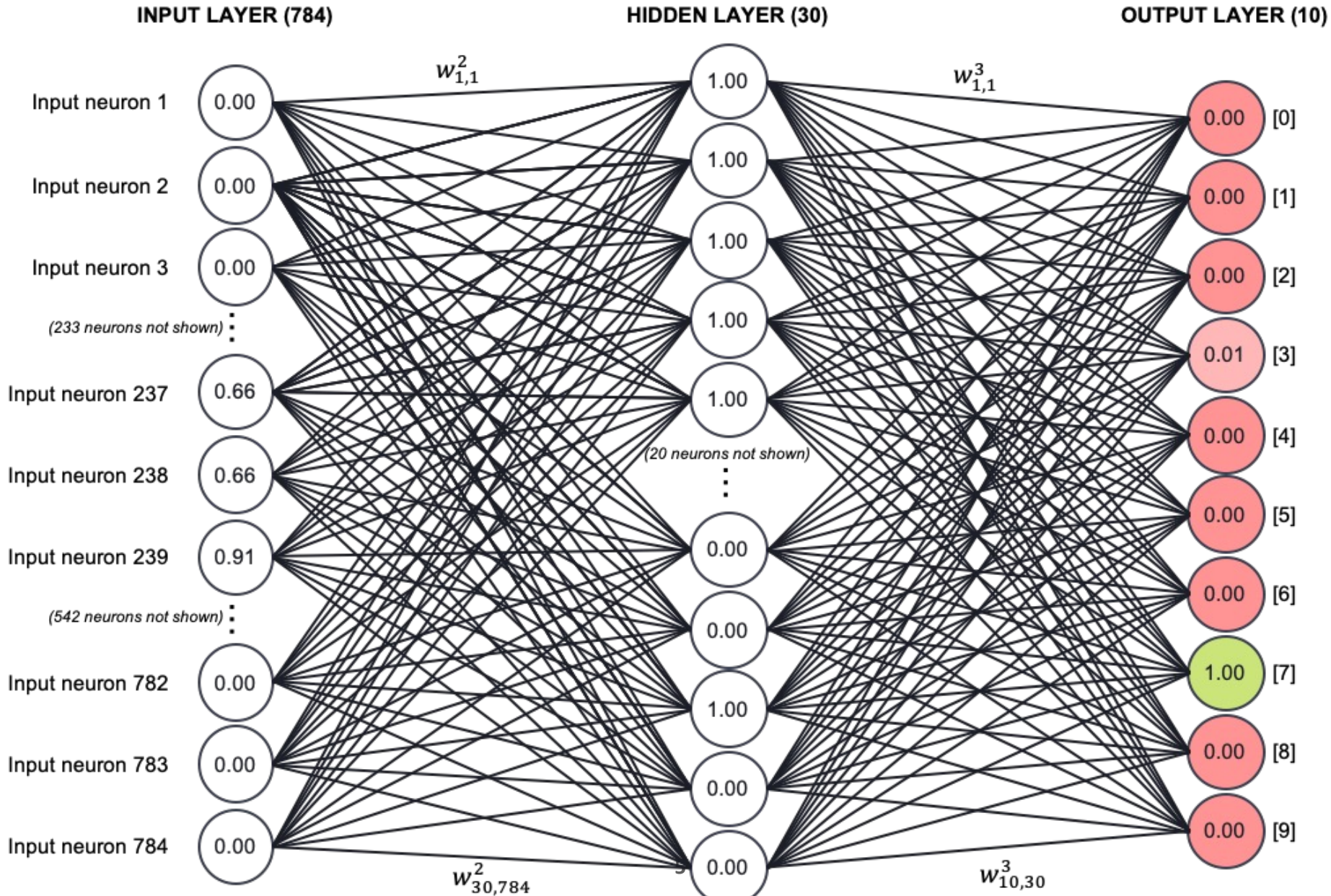
Goal

- This project provides various **exploratory methods** to
  - analyze any NER model/dataset combination,
  - find labeling errors,
  - understand the model's and dataset's limitations.
- Exemplified with a DistilBERT model [4] + ConLL03 dataset [5]

Named Entity Recognition



Neural Networks: Vectors and Matrices



Dataset: CoNLL03 [5]

Data format

U.N.		I-ORG
official		O
Ekeus		I-PER
heads		O
for		O
Baghdad		I-LOC
.		O

Dataset size

English data	Articles	Sentences	Tokens
Training set	946	14,987	203,621
Development set	216	3,466	51,362
Test set	231	3,684	46,435

Label distribution

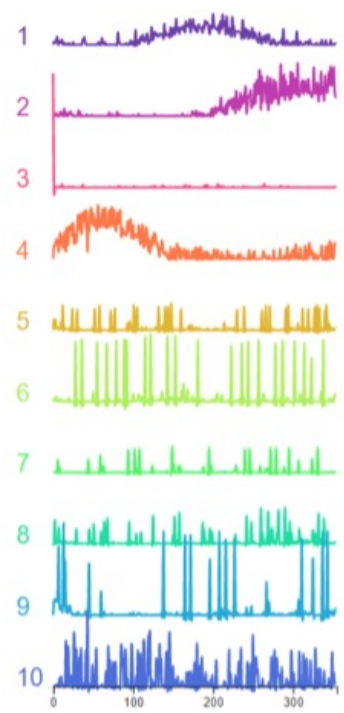
English data	LOC	MISC	ORG	PER
Training set	7140	3438	6321	6600
Development set	1837	922	1341	1842
Test set	1668	702	1661	1617

Implementations

- Interactive **sparkline** visualization of the **neural network activations**
- Interactive **similarity map** of (a 2d-projection of) the **model's final layer's hidden states** (which are used for token classification)
- **HTML representation** of **dataset samples** with token-level prediction + labels (extremely information-dense but highly useful)

Neuron Activation Visualization (from Alammar, 2021 [1])

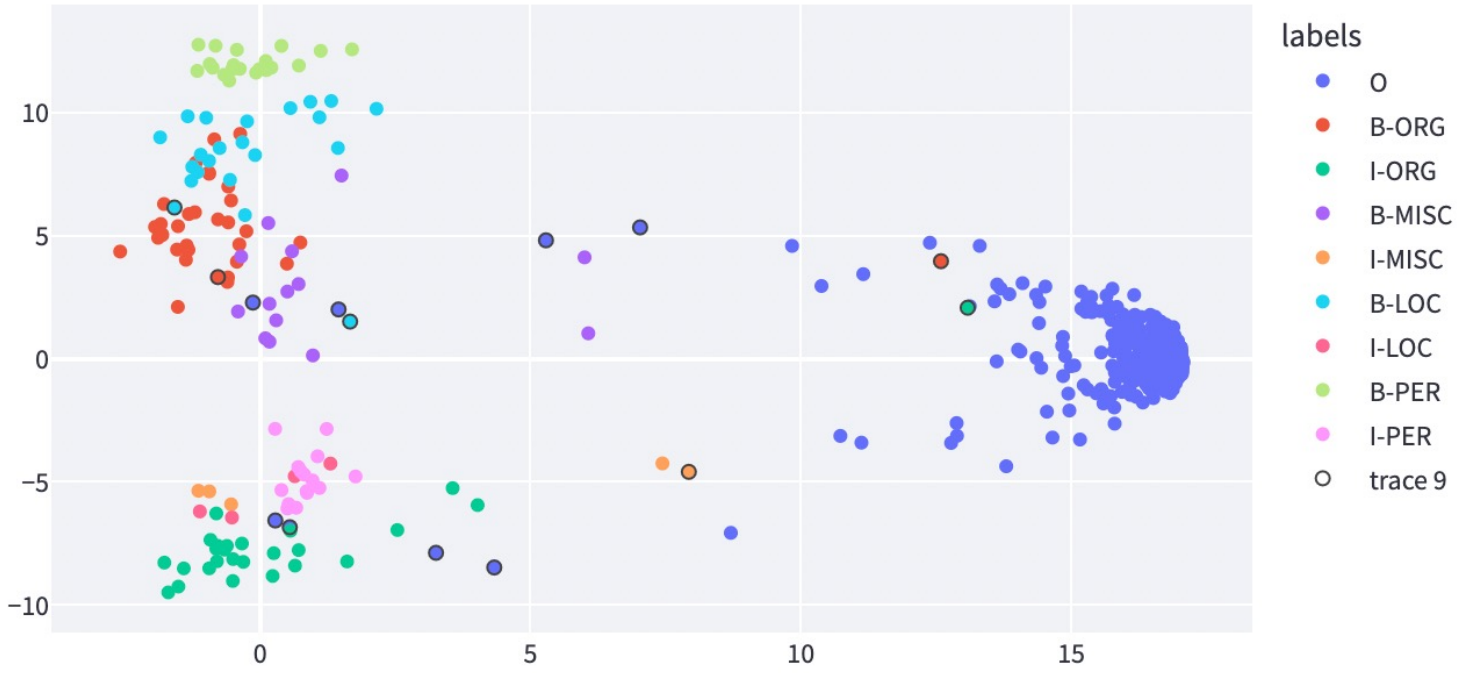
- Inner workings of **neural networks** are **still poorly understood**
- Non-negative Matrix Factorization reveals **underlying patterns of neuron activations** inside the model's layers
- **Interesting result:** We can see that different factors correspond to different textual properties: pronouns, punctuation, beginning, middle, end, etc.



Now I ask you : what can be expected of man since he is a being endowed with strange qualities ? Sh ower upon him every earthly blessing , drown him in a sea of happiness , so that nothing but bubbles of bliss can be seen on the surface ; give him economic prosperity , such that he should have nothing else to do but sleep , eat cakes and busy himself with the continuation of his species , and even then out of sheer ing rat itude , sheer spite , man would play you some nasty trick . He would even risk his cakes and would deliberately desire the most fatal rubbish , the most un econom ical absurdity , simply to introduce into all this positive good sense his fatal fantastic element . It is just his fantastic dreams , his vulgar folly that he will desire to retain , simply in order to prove to himself -- as though that were so necessary -- that men still are men and not the keys of a piano , which the laws of nature threaten to control so completely that soon one will be able to desire nothing but by the calendar . And that is not all : even if man really were nothing but a piano - key , even if this were proved to him by natural science and mathematics , even then he would not become reasonable , but would purposely do something perverse out of simple ing rat itude , simply to gain his point . And if he does not find means he will cont rive destruction and chaos , will cont rive suffer ings of all sorts , only to gain his point ! He will launch a curse upon the world , and as only man can curse (it is his privilege , the primary distinction between him and other animals), may be by his curse alone he will attain his object -- that is , convince himself that he is a man and not a piano - key ! \n

Similarity Map (like TensorFlow Embedding Projector [7])

- **Background:** Texts are represented as **high-dimensional vectors**
- **Idea:** Reduce dimensionality of those vectors so we can plot them **onto a two-dimensional plane**
- **Benefit:** By coloring data points by label/prediction, with disagreements marked by a small black border, we can visually inspect the dataset to **find mislabeled examples**.



Libraries

- **streamlit** for demoing
- **plotly** and **matplotlib** for charting
- **transformers** for providing the models, and **datasets** for, well, the datasets
- a forked, slightly modified version of **ecco** for visualizing the neural net activations
- **sentence_transformers** for finding potential duplicates
- **scikit-learn** for TruncatedSVD & PCA, **umap-learn** for UMAP

Main Paper

Alammar, Jay

Ecco: An Open Source Library for the Explainability of Transformer Language Models

Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations, pages 249–257, August 1st – August 6th, 2021.

References

- [1] Alammar, J. (2021). Ecco: An open source library for the explainability of transformer language models. https://doi.org/10.18653/v1/2021.acl-demo.30
- [2] Data Centric AI Workshop, NeurIPS 2021, https://neurips.cc/Conferences/2021/ScheduleMultitrack?event=21860
- [3] Northcutt, C. G., Athalye, A., & Mueller, J. (2021). *Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks*. *NeurIPS*. http://arxiv.org/abs/2103.14749
- [4] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*. http://arxiv.org/abs/1910.01108
- [5] Tjong Kim Sang, E. F., & De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task. https://doi.org/10.3115/1119176.1119195
- [6] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. http://arxiv.org/abs/*1706.03762*
- [7] TensorFlow Embedding Projector, https://projector.tensorflow.org/