NxPlain

A Web-based Tool for Discovery of Latent Concepts

Fahim Dalvi Nadir Durrani Hassan Sajjad
Tamim Jabban Mus'ab Husaini Ummar Abbas
{faimaduddin,ndurrani}@hbku.edu.qa

Background

Interpretation of Deep Learning Models can be broadly divided into two branches

Representation Analysis
What is learned by a model?

Attribution Analysis
What is used by a model to predict?

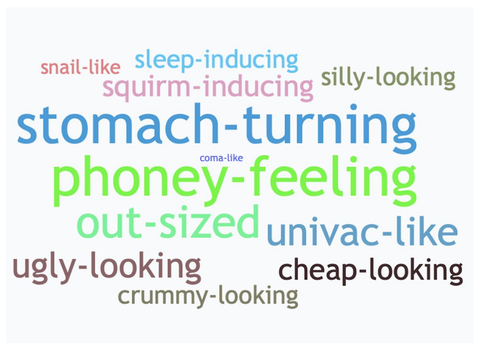
Goal

Provide a holistic view into the model by combining representation and attribution analysis

Overview

Leverage *Latent Concept Analysis* [1], *Concept Alignment* [2] and *Attribution Analysis* [3] to explain *what knowledge* a model has learned and *how does it use the knowledge* to make predictions

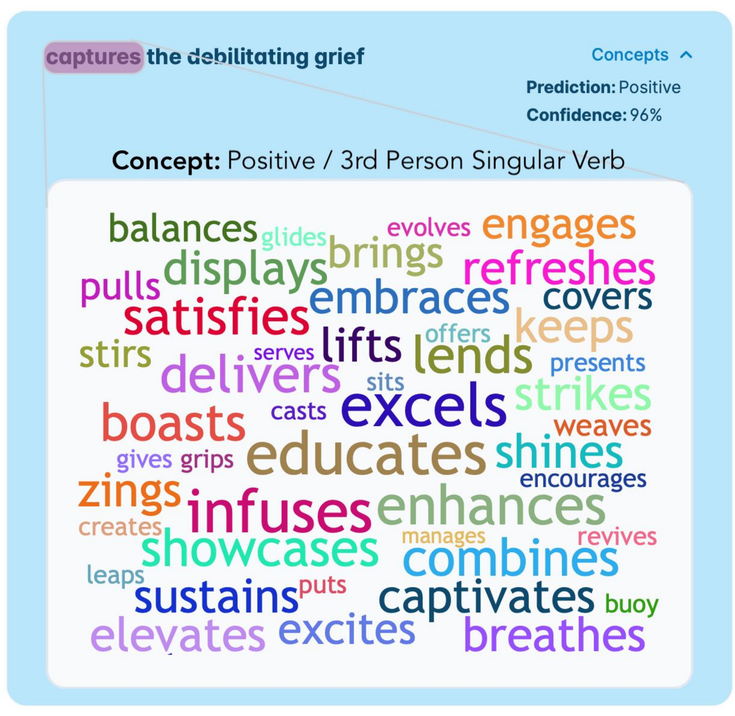
Terms used in hate-speech against immigration policies



Syntactic concept of hyphenated words

Concept made up of numbers

Named entities in Germany

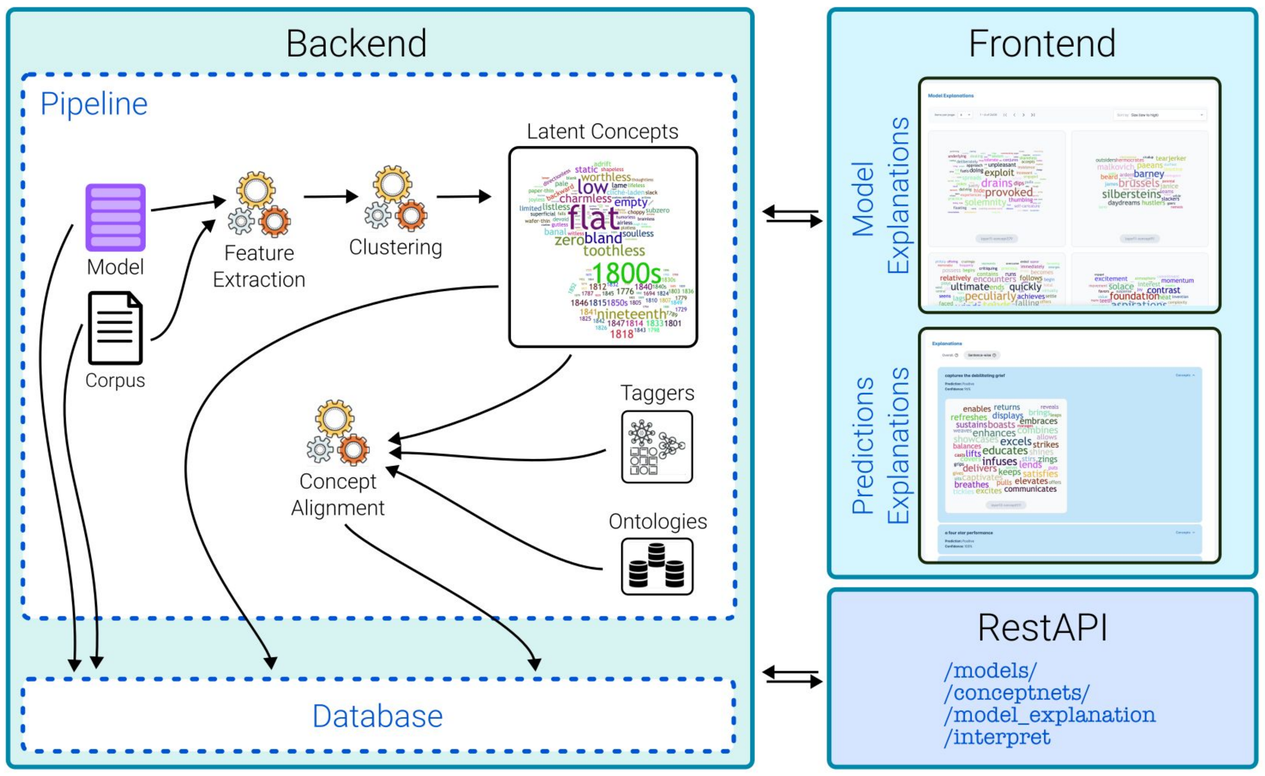


[1] Dalvi, Fahim, et al. "Discovering Latent Concepts Learned in BERT." International Conference on Learning Representations 2022

[2] Sajjad, Hassan, et al. "Analyzing Encoded Concepts in Transformer Language Models." North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2022

[3] Sundararajan, Mukund, Ankur Taly, and Qiqi Yan. "Axiomatic attribution for deep networks." International conference on machine learning. PMLR, 2017

Architecture



Try it out

https://nxplain.qcri.org

Start understanding your models today!

NeuroX
https://neurox.qcri.org

QCRI
معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY