ConEm: Learning Embedded Concept Representations from LLMs

Luna Peck, Alvin Chen, Maria Pacheco (University of Colorado Boulder)

Boulder Language And Social Technologies

Research Question

Can we learn non-linguistic representations that "make sense" to language models?

ConEm: Concept Embeddings

- Capture **general concepts** like "positive sentiment," "review," "customer experience," without relying on specific natural language representations.
- **Learned directly from the model** by next-word-prediction on texts that are examples of relevant concepts.
- Can be **composed into modular prompts** indicating multiple concepts, such as the context of a text.
- Potential use as **predictable inputs** to language models, symbolic representations in **neuro-symbolic systems**, or representations for studying **neurally-encoded knowledge.**

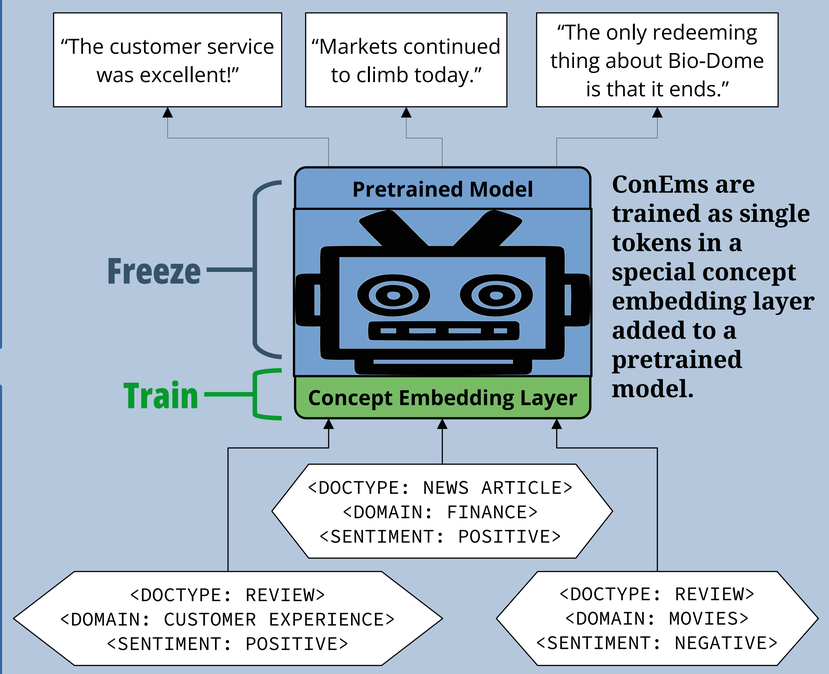
Evaluation: Contextualized Generative Classification

Generative classification reverses the usual order of classification: class labels are used as inputs across separate inferences, and the class label that has the highest probability of producing the text-to-classify as output is chosen as the predicted label.

Experiments adding context to generative classification input have shown improved performance. We compare zero-shot contextualized generative classification using natural language sentences to contextualized generative classification using trained ConEms.

Motivation

Language models are trained on linguistic representations, but it is often difficult to know which linguistic representation is best for a given task. Models are **sensitive to minor changes in wording** that seem inconsequential to humans, and **ideal linguistic representations are often unintuitive**.



ConEms are trained as single tokens in a special concept embedding layer added to a pretrained model.

ConEms are typed, allowing directly comparable prompts to be constructed by placing ConEms of the appropriate type into ordered slots. This is helpful for comparing sentiment ConEms in the generative classification task.

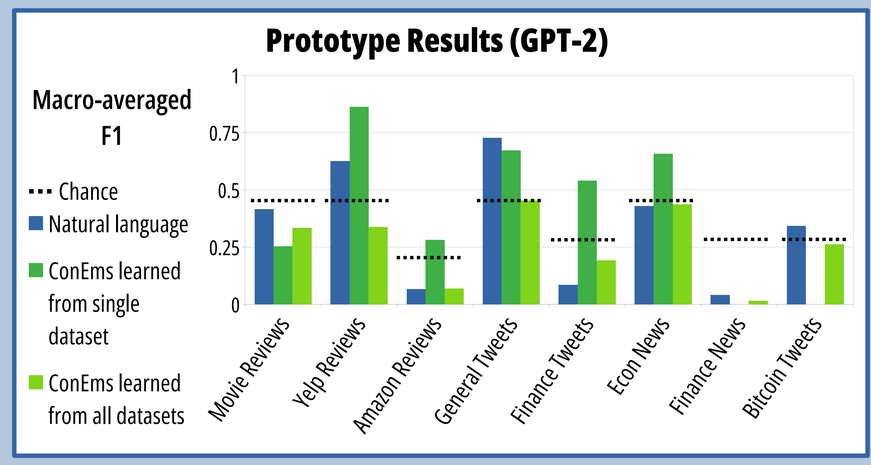
Some ConEms were trained only on the training portion of the dataset on which they were tested. Others were trained on the training portions of all non-held-out datasets.

Datasets used:

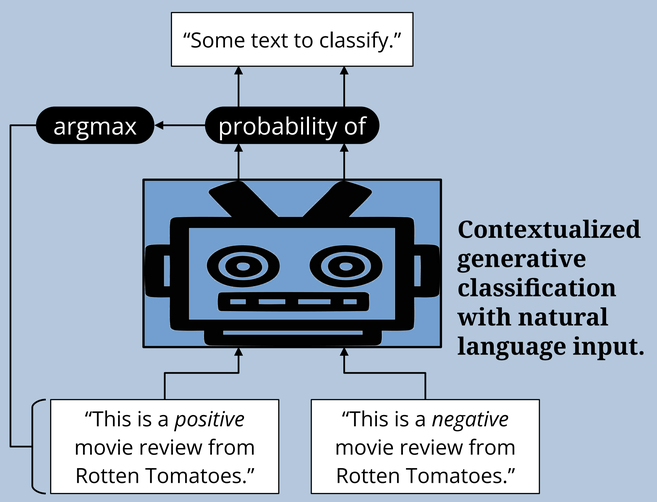
Dataset	Context ConEms
Movie Reviews	**<Origin:** Rotten Tomatoes> **<Domain:** Film> **<Doctype:** Review> **<Sentiment:** {Negative, Positive}>
Yelp Reviews	**<Origin:** Yelp> **<Domain:** Consumer Business> **<Doctype:** Review> **<Sentiment:** {Negative, Positive}>
Amazon Reviews	**<Origin:** Amazon> **<Domain:** Consumer Products> **<Doctype:** Review> **<Sentiment:** {Negative, Negative+Neutral, Neutral, Positive+Neutral, Positive}>
General Tweets	**<Origin:** Twitter> **<Sentiment:** {Negative, Positive}>
Finance Tweets	**<Origin:** Twitter> **<Domain:** Finance> **<Doctype:** News> **<Sentiment:** {Negative, Neutral, Positive}>
Econ News	**<Origin:** Mainstream News> **<Domain:** Economics> **<Doctype:** News> **<Sentiment:** {Negative, Positive}>
Finance News (held out)	**<Origin:** {Mainstream News, Twitter}> **<Domain:** Finance> <Doctype: News> **<Sentiment:** {Negative, Neutral, Positive}>
Bitcoin Tweets (held out)	**<Origin:** Twitter> **<Domain:** Finance> **<Doctype:** News> **<Sentiment:** {Negative, Neutral, Positive}>

Some datasets are held out from training to test the generalizability and compositionality of ConEms.

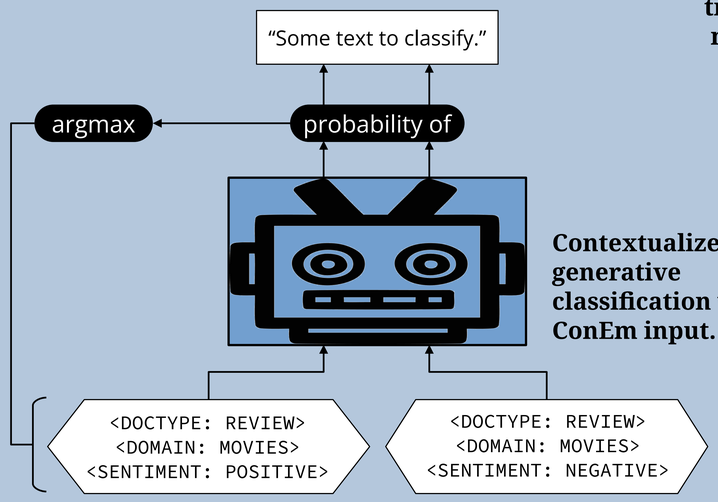
Prototype Results (GPT-2)



Note: No single-dataset-trained ConEms for held out datsets.



Contextualized generative classification with natural language input.



Contextualized generative classification with ConEm input.

Discussion & Next steps

Results are inconsistent, with highly varying performance across tasks and no consistent trend across methods. However, **all datasets** except Movie Reviews and Finance News saw **above-chance performance for single-dataset-trained ConEms**, suggesting the method is worth pursuing further. We will investigate the MR and FN datasets for properties that may explain poor ConEm performance. We will also investigate performance on larger, modern language models.

We are currently rewriting the ConEm training loop from scratch, to give us finer control over the training process. We are also investigating our generative classification implementation, as we believe we can improve it.