3-20-2024

Natural Language Processing: Understanding Slang and Colloquial Speech

Beau Miller

Natural Language Processing: Understanding Slang and Colloquial Speech

BEAU MILLER

Research Symposium Presentation

March 20th, 2024

Abstract

Within the field of Natural Language Processing, slang and colloquial language provide a unique challenge for the training and use of language models. This speech holds an odd, transient space with public consciousness, as each instance has a variable time of relevancy and popularity, amongst other factors such as their fluid definition, usage amongst certain groups, and relative lack of information and training data concerning new or evolving terms. These characteristics make it difficult for models to learn these phrases in a timely manner.

This project focuses on how a NLP model could learn new additions to colloquial language in a timely and efficient manner, as well as keep up with a rapidly mutating lexicon.

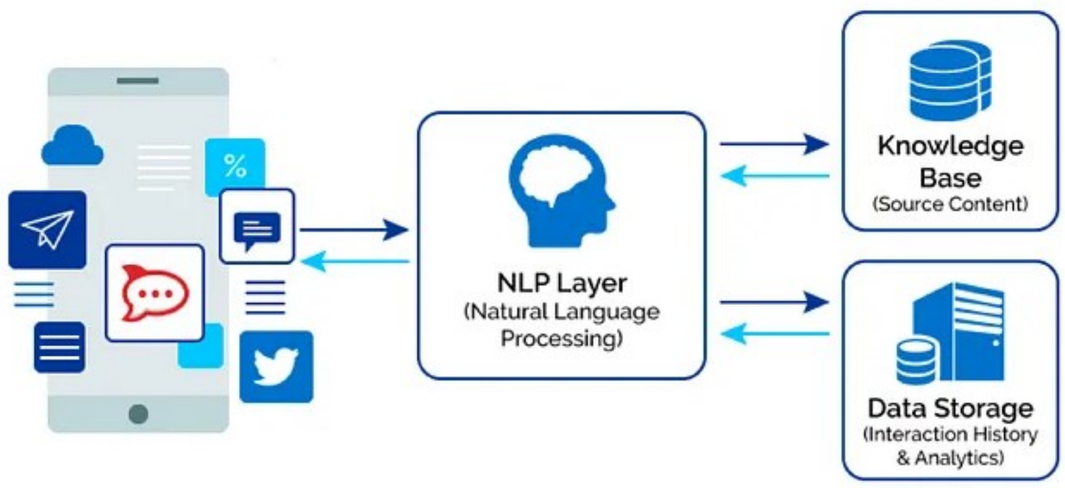

Figure 1: A chart[1] demonstrating a simplified NLP model architecture.

Background

Within language, slang and colloquial speech are an important subset; these informal words and phrases are both a comfortable shorthand for communication and act as a reflection of a culture, whether it be universal, ethnic, or local. Many people use both Slang and Colloquial Speech within their everyday speech, often without noticing. This speech may range from the benign, occasional use of a common phrase to the heavy usage of more obscure or new colloquial phrases.

While NLP models are capable of incorporating these colloquial phrases into their lexicon, the acquisition of new phrases and updating old phrases takes time for a model to learn, and in addition, the data required to train the models may prove limited and subpar. For more universal, long lasting slang such as the term ‘cool’, this may be acceptable, but the problem begins to manifest for lesser known and volatile slang terminology, whose available data may not prove sufficient for training and fine tuning a model.

The ability for a NLP model to quickly understand and adapt to the shifting landscape of slang is important for universal accessibility and a better ability to adapt to an ever-shifting culture and vernacular.

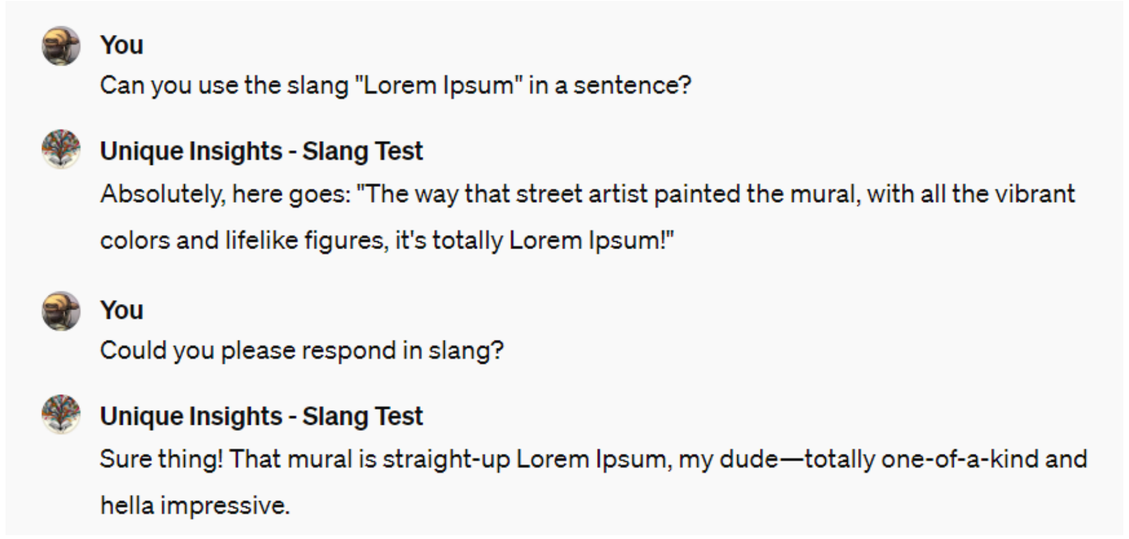

Figure 2: a custom GPT model designed to utilize the made up slang term ‘lorem ipsum’.

Current Findings

As demonstrated in *Figure 2*, while it is simple enough to instruct a language model to use a word with a modified meaning, properly integrating a phrase into a model takes much more work. It is not as simple as slotting in a new word and a definition; the model must also know the context in which to use the word, how to use it in a sentence, the potential nuance the word possesses, amongst other factors. Hence, why it takes time and plenty of training data to properly teach a model a new word and its various intricacies.

The main bottleneck for training NLP models on slang and colloquial phrases is the availability of data to train a model on, as well as the quality and context of the data itself. Fine tuning the pre-existing model with said data, conversely, is a relatively quick process.

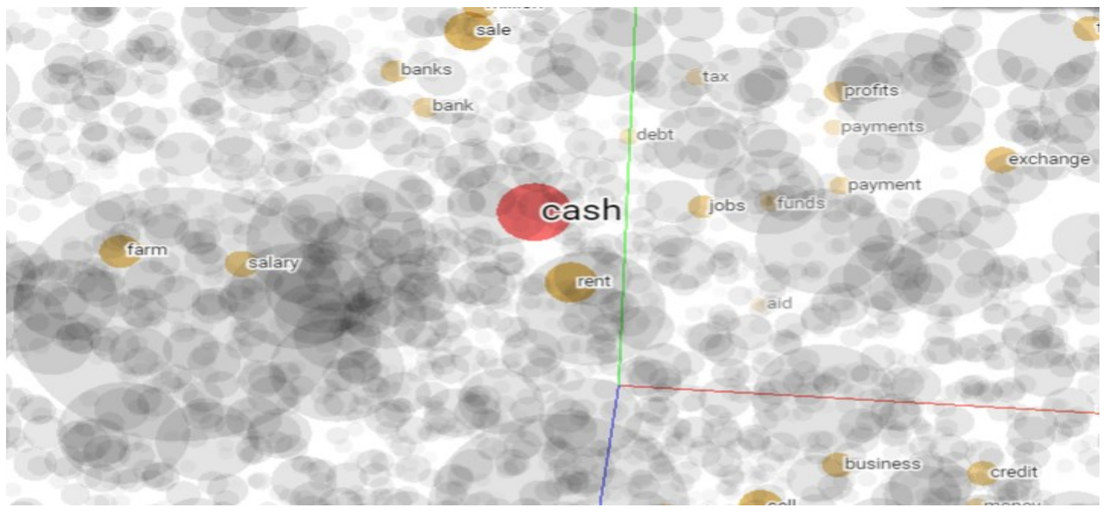

Figure 3: A 3D model representing an embedding of words utilized by some NLP models. In this specific example, the term ‘cash’ is selected, and highlighted are other terms it is related to.

Conclusions

Natural Language Processing and how it interacts with Slang and Colloquial Speech is an important facet to consider within this evolving field. As such speech is inseparable from humanity in an everyday, casual environment, the ability to adapt it into an efficient NLP model proves important for its ultimate efficacy, especially for practical applications concerning public everyday interactions and a greater integration and understanding of the cultural landscape.

[1] S. Kayte, "Natural Language Principles," Medium, December 5, 2020. [Online]. Available: “https://chatbotslife.com/natural-language-principles-65e88e20b94”