ML Semantic Analysis of Tone Indicators on Twitter

Eigen Ott, Run Chen, Rohan Sahu, Charlie Jiang, Shivansh Srivastava

Introduction

Tone indicators (written as /tone) are a tool used to **reduce ambiguity**, conveying information that is generally lost online. The literal text of informal communication like a tweet is often **insufficient to determine the meaning**, which relies on much a larger, inaccessible context. Tone indicators attempt to address this issue.

Using data from Twitter, we built a BERT model to **classify the tone indicator tag** associated with a sentence. We then analyzed the various **attention heads** to gain semantic insight into which sets of words or other linguistic features are most associated with certain tone indicators.

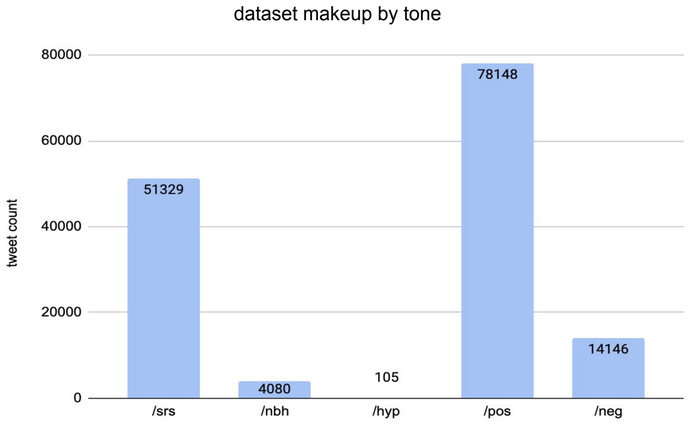
Many tone indicators are popularly used in certain online communities, but we focused on the five shown below.

tag	/srs	/nbh	/hyp	/pos	/neg
meaning	serious	nobody here	hyperbole	positive	negative
usage	when being serious (ie not joking or sarcastic)	talking about someone else not present (generally negative)	to note exaggerated speech	to signify positive emotions	to signify negative emotions
examples from the dataset	Can someone dm me with what's going on? /srs	shut up shut up shut up oh my fucking gods stfu /nbh	ohhhh my god if i don't get faster/better at LaTeX soon i will die /hyp	crying and sobbing so hard right now /pos	wasn't able to go to school for 4 days last week bc my mental health, i love being me /neg

Dataset

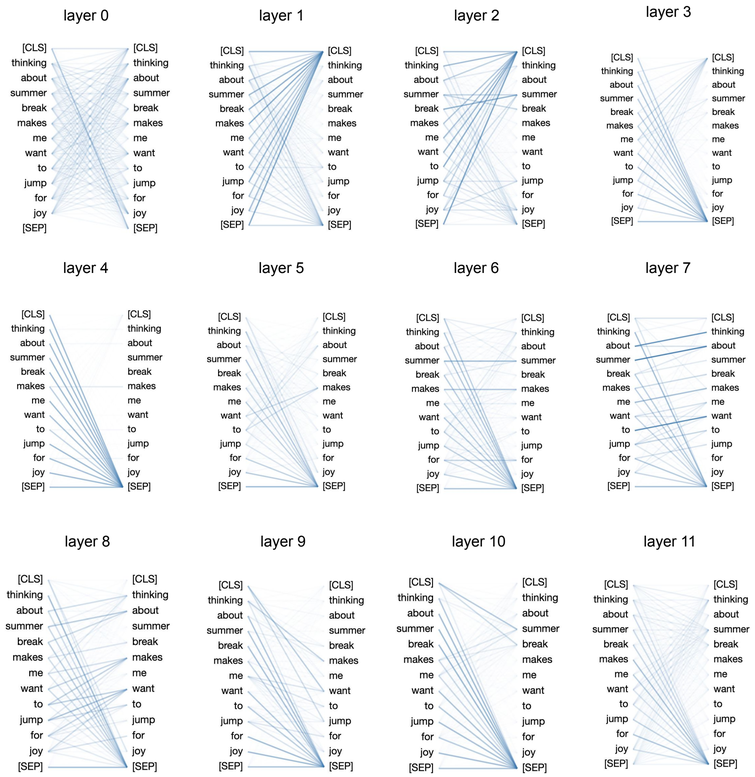
Using the Twitter API, we scraped over 1.5 million tweets between January and March of 2022. The language of each tweet was determined using langdetect, which also cleans excess emojis. The end result was a subset of approximately **150,000 tweets**.

We also obtained a training set consisting of a separate and smaller scrape of tweets from 3 random days in February of 2022. All data in the training set was labeled with the emotional tag that it contained using one-hot encoding. All tags were stripped out and replaced with white space in the test set.

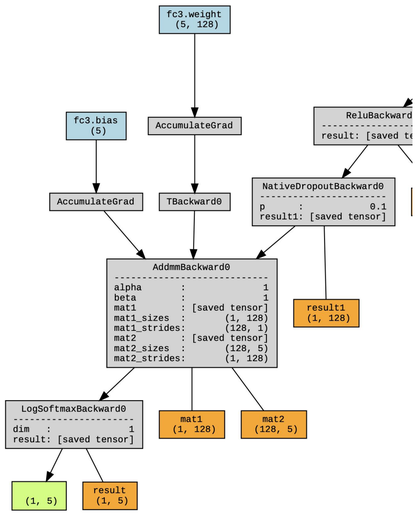


Attention and Architecture

Self-attention across all model layers



Final BERT Model Architecture Layer



- Pre-existing BERT architecture fine-tuned for tone indicator classification
- 12 total layers
- Added custom layers for this task
- 3 linear layers following Bert architecture
- Produces (1,5) vector output
- Feed in bias and weight

Transformers

vinai/bertweet-base like 8

Fill-Mask PyTorch TensorFlow JAX Transformers

Results

Model	Accuracy
Naive Bayes	70.02
BERT Model	**97.76**

The BERT model significantly outperformed the Naive Bayes classifier. Notably, the Naive Bayes only had three categorization options (pos, neg, and a third "neutral" category of hyp, nbh, and srs).

We used /vinai/bertweet-base which is specifically trained on tweets as the base model before fine-tuning, and from the attention visualizations at left, it's clear that the BERT model picked up on additional linguistic features more complex than just specific tone-signifying words.

Semantic Word-Tone Associations

	Higher freq →								Lower freq	
/srs	pls	drop	actually	?	guys	tweet	any	anything		
/nbh	stop	ppl	yall	then	person	bc	you're	youre	talk	off
/hyp	=	er	ever	myself	/hj	The	has	joke	question	than
/pos	cry	SO	crying	:(	THE	him	art	AND	IS	THIS
/neg	//	also	had							

We found the 100 most frequent words for each tag. In order to isolate the words uniquely associated with a tag, we performed **set difference operations** between all 5 sets to eliminate overlap.

The above were revealed to be the top 10 most common unique words for each tag. These words are heavily influenced by internet slang and culture.

Conclusion & Future Work

There is a consistent style of usage to tone indicators that can be learned by a machine learning model to a high degree (human level) of accuracy. Tone indicators can be treated a novel feature of of natural language with unique linguistic rules. Since we only studied five tags, future work would naturally extend to covering the rest of the tone indicator tags that are commonly used. Additionally, due to constraints on tweets (140 characters max) more varied datasets could yield more robust results.

References

[1] What Does BERT Look At? An Analysis of BERT's Attention
Kevin Clark, Urvashi Khandelwal, Omer Levy, Christopher D. Manning
https://arxiv.org/abs/1906.04341

[2] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova
https://arxiv.org/abs/1810.04805

[3] Tone indicators: what they are, why you should use them, and how to use them?
https://toneindicators.carrd.co/

[4] BertViz: Visualize attention in NLP Models. Jesse Vig, Martin Sotir, Phillip Glock
https://github.com/jessevig/bertviz