Uncovering the Causes of Emotions in Software Developer Communication Using Zero-shot LLMs

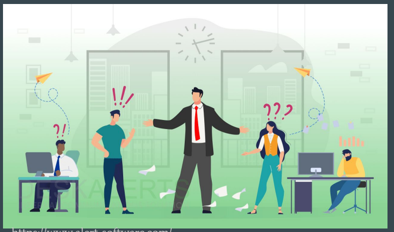
By Mia Mohammad Imran, Preetha Chatterjee, and Kostadin Damevski

•••

Presented by Daniel Pang
2025/01/23

The Problem

- Understanding and identifying causes of developer emotions is difficult
  - Especially across multiple communication channels (chats, emails, comments, etc.)
- LLMs are a good solution, commonly used (e.x. Twitter)
- Training is expensive, curating a dataset for this use case is expensive



https://www.alert-software.com/

The Idea

- Test efficacy of zero-shot LLMs to detect emotional causes in software engineering
  - Zero-shot LLMs being LLMs that have been trained on massive dataset but not specifically for a specific task (i.e. this task)
  - This paper tests ChatGPT, GPT-4, and flan-alpaca on three datasets
- Compares performance to state-of-the art techniques
  - Existing SE-specific models for emotion classification
  - Fine-tune LLMs that are already good at emotion classification
- Perform a case study
  - Finding sources of "Frustration" in open source project Tensorflow

TensorFlow

https://www.tensorflow.org/extras/tensorflow_brand_guidelines.pdf

Quick Results in Emotion Classification

- The zero-shots did relatively ok
  - Predicted conservatively (neutral)
  - Hallucination is a problem
    - Hallucinations get worse with more emotions
  - Really uneven distribution

	Imran et al. (N=2000) [21]						
	Anger	Love	Fear	Joy	Sad.	Surprise	Micro Avg.
SE-Specific							
ESEM-E	0.309	0.644	0.291	0.378	0.524	0.500	0.440
EMTk	0.430	0.682	0.163	0.378	0.525	0.345	0.434
SEntiMoji	0.460	0.642	0.377	0.556	0.629	0.458	0.529
Fine-tuned LLMs							
BERT	0.506	**0.712**	0.536	**0.579**	**0.636**	0.594	0.588
RoBERTa	0.525	0.683	0.492	0.500	0.613	0.673	**0.592**
Zero-shot LLMs							
ChatGPT	0.337	0.492	0.182	0.458	0.417	0.511	0.429
flan-alpaca	0.447	0.537	0.140	0.446	0.451	**0.740**	0.506
GPT-4	0.409	0.698	0.049	0.446	0.487	0.524	0.482

Quick Results in Cause Extraction

- The authors seems pretty happy with the BLEU scores
    - Only 41/750 cases had all 3 models have a score <0.30 (garbage response)
- Biggest issue is misclassification of emotion leading to misidentification of cause
- Sometimes emotion is correct but cause is incorrect
- As usual, sometimes LLMs hallucinate and return nonsense

Table 4: BLEU scores of different zero-shot LLMs.

Model	BLEU-1	*BLEU-2*	BLEU-3	BLEU-4
ChatGPT	0.522	*0.489*	0.467	0.450
GPT-4	**0.637**	***0.598***	**0.571**	**0.554**
flan-alpaca	0.571	*0.543*	0.525	0.508

Quick Results on the Case Study

- Utilized flan-alpaca as it did the best in their tests
    - Also open source for reproducibility
- Took almost all comments from March 2022 to March 2023
- Clustered using DBSCAN, with manual testing of different parameters
- Results are good

Table 5: Clusters of causes of *Frustration* in TensorFlow project participants in GitHub.

	Cluster Description	Count	Example Comments
1	**TensorFlow Version and Dependency Issues**: This cluster focuses on build and compatibility problems across various TensorFlow versions, challenges in reproducing issues in specific TensorFlow versions, and complications with related libraries and plugins such as TensorRT and Keras.	58	(1) *[USER] Your original issue looks like* ***you have a bad version of tensorflow_io_gcs_filesystem installed****. [...].* (2) ***It's probably not a bug in Tensorflow but Apple's tensorflow metal plugin****. See for example the following discussion [...]*
2	**Pull Request Delays and Merge Conflicts**: The cluster comprises developer frustration from unresolved merge conflicts and from delays in merging pull requests.	26	(1) *[...] But* ***there are a bunch of merge conflicts****. Since Random seeds are such a common topic in software [...]* (2) *It might have been a* ***wrong-way merge*** *or something like that. At this point it's usually easier to just close it [...]*
3	**Failing Tests**: This type of *Frustration* arises from the ambiguity and complexity of test failures, which make it challenging for project participants to determine whether the issues are linked to their code changes or are caused by unrelated factors.	15	(1) *[USER]: It is just a first draft.* ***The test doesn't even work****. In the meantime, [...]* (2) *[...] Yes, I'll work on this.* ***It's weird that these tests are failing because I thought I ran them successfully for PR*** *[...]*
4	**Too Fine-Grained Commits**: The cluster reflects developer *Frustration* caused by too granular commits in the repository. Some developers request a commit history devoid of incremental commits that represent only partial progress on a change task.	9	(1) *Can you squash these commits please?* ***It doesn't make sense to have 5 commits for a line change*** *and one extra empty line.* (2) ***3 commits for a single line change****? Can you please merge the commits in just one? [...]*
5	**CI Flakiness**: This type of *Frustration* is caused by Continuous Integration (CI) failures that seem unrelated, inconsistent, or uninformative to developers.	8	(1) *[USER]* ***there was failed ci.*** *Is there anything to do?* (2) ***CI failure does not look related to these changes****, seeing the same failure on #56345 [...] so I assume this is noise. [...]*
6	**CUDA/CuDNN Compatibility Issues**: This cluster reflects the *Frustration* experienced when dealing with compatibility issues related to CUDA and CuDNN.	8	(1) *Unfortunately this change needs to be rolled back,* ***it seems it breaks JAX build under CUDA 11.4 and CuDNN 8.2*** (2) *[...] - Did you downgrade the CUDA to 11.2?* ***Looking at Nvidia docs it looks like the display driver and cuda driver do not match*** *[...]*

Positives

- Little research into the performance of zero-shot LLMs for SE
    - Seeing a notable gap between zero-shot and SE-specific or Fine-tuned LLMs illustrates the complexity of SE communication
    - However, although it's a notable gap it's not a massive gap, it's a solid option for low cost
- Really great use case
    - Even if emotion cause detection is not consistent, you don't need perfection to get a good overview of the problems
- A lot of various testing
    - 3 different emotional models
    - 8 different LLMs
    - 3 different datasets

Negatives

- Emotions are complicated and categorization is imperfect
    - Some of the test failures that they showed as examples illustrate the ambiguity of the English language and show that the LLMs are close but do not provide the exact correct answer
    - E.x. "Ah sorry I thought 'ScaleUpdateDetails' was constructed in '_update' nvm" was annotated as "Sadness" but predicted with "Surprise"
- LLMs are incredibly ambiguous
    - How their fine-tuned LLMs are tuned are unknown
    - For the existing SE-specific model ESEM-E, the authors had to implement the model themselves from instructions, adding another layer of ambiguity
    - Reproducibility is difficult

Future Work

- Several mentioned
    - Future case studies can work on multiple emotions and/or multiple projects
    - Improvement on extracting causes of emotions from text in SE
- Personally
    - Improvement on categorization of emotions for LLM recognition
        - Realistically statements can have multiple emotions
        - Even if the data set only has one correct emotion for each utterance, it is sometimes very hard to draw the line between two emotions

Rating

4

I think the use case is very useful and the research into the efficacy is pretty good. However, both LLMs and human emotion are two things that are very ambiguous and difficult to define, leading to a feeling of the paper being a little nebulous

Discussion Points

- Would you want your organization to use LLMs to attempt to get a read on emotions and/or find causes of emotions?

- Comparison of finding sources of certain emotions to the work of card sorting

- Where would this fit along the spectrum of data collection from questionnaires to interviews?