CS 846 - Causes of Emotions Using Zero-shot LLMs

Daniel Pang
daniel.pang@uwaterloo.ca

ACM Reference Format:
Daniel Pang. 2025. CS 846 - Causes of Emotions Using Zero-shot LLMs. In . ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 PROBLEM BEING SOLVED

The authors present the issue of trying to identify the causes of what is causing specific emotions across the many different channels of communication for any project, including chats, emails, issue comments, etc. Using LLMs would be a good idea, however training as well as curating a dataset to train on are both cost intensive.

2 NEW IDEA

The authors explore the efficacy of using zero-shot LLMs, aka LLMs that have not bee trained for a specific purpose, for this task, thus eliminating a large cost at the detriment of specialization. To analyze how well these zero-shot LLMs perform, three zero-shot LLMs are compared to emotion classification models that are SE-specific and fine-tuning emotion classification LLMs to SE in particular.

The authors found that the zero-shots did relatively well in contrast to the competition in categorizing emotions, and although they did notably worse, it wasn't too much worse. However, there were some issues that the authors outlined in that the zero-shots tended to predict conservatively and tended to hallucinate when given more emotion categories.

Next was the test to attempt to extract the cause of the emotion in a given utterance. Again, according to the authors, the zero-shot LLMs did reasonably well when utilizing BLEU scores comparing the returned string to hand-annotated correct answers. In fact, upon error analysis, the authors found that the biggest contributor to incorrect conclusions was simply a misclassification of the emotion of the utterance, which stems from the previous step.

Finally, the authors conducted their own case study using flan-alpaca to extract causes of Frustration from the open source project Tensorflow. The results seem to be good as well here as they managed to extract from a years worth of comments to find sources of frustration that were nicely clustered into several categories.

3 POSITIVE POINTS

Firstly, there is little research into the performance of zero-shot LLMs for SE and I think this should be further explored as, as the authors mentioned, it is very costly to procure a dataset as well as train an LLM. Further, the results of this paper show a reasonably small gap between zero-shot LLMs and common alternatives.

Secondly, I think the use case is very useful. Even if emotion cause detection is not the most consistent, getting a general feel for causes of emotions provides a great direction for further investigation.

Lastly, I was happy with how much testing went into this paper. Notably there were 3 different emotional models tested using 8 different models (LLMs + ML models) over 3 different datasets. This allowed the authors to illustrate some nice differences as well as commonalities between the different options.

4 NEGATIVE POINTS

Firstly, emotions are incredibly complicated and I found that categorization is both very difficult and imperfect. Notably, some of the test failures that were provided in the paper I found illustrated the ambiguity of the English language, especially when written down as text. In these cases the conclusion the LLM reached would be close but would not exactly match the annotated correct answer

Secondly, LLMs are incredibly ambiguous, they are a black box that makes making conclusions difficult. Further how the authors fine-tuned the LLMs to compare to is unknown and in addition for the existing SE-specific model ESEM-E, the authors had to implement the model themselves which is another failure point. All this leads to difficulties in reproducibility.

5 FUTURE WORK

They were several potential future works mentioned in the paper themselves: future case studies working on multiple emotions and/or multiple projects and improvement on the extraction process. Personally I had issues with categorization on emotions and would like to see that improved. This is because I find emotions to be very nuanced, for example it can be very difficult to distinguish the emotion from a given statement even as a human reader.

6 RATING

I still give the paper a 4 despite the flaws. This is because I highly rate the use case and found the research and testing conducted by the paper to be quite nice. However, LLMs and human emotion are both very ambiguous and difficult to defined, leading to feeling of the paper being nebulous.

7 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 interview?

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

8 IN-CLASS DISCUSSION

Firstly, the validity of BLEU scores was brought up to discredit the efficacy of the LLMs cause extraction. This is because BLEU scores are used to compare machine translations to human translations across languages and apparently even in that case their validity is questionable. However, I find that the usage here is perfectly valid as the usage is simply to compare how similar the LLMs conclusions are to the annotated answer that was provided by undergraduate students. Simply it is comparing how similar two strings of text are, the prediction to the correct answer. Further, if that were not enough, the results of the case study still show a success despite this, as they were able to extract causes from comments and cluster the results well into logical groupings.

Extending on the previous was a discussion on what should be done with a paper if methodology may be flawed where I answered that a follow up study could be performed to test reproducibility or just as a critique of the paper. I believe there are even meta studies that do this exact thing.

A student had then brought up an idea of seeing a follow up study using a real privatized company instead of an OSS. They mention stresses, such as deadlines, could skew the sorts of emotions present in communication channels to which I agreed with examples like a slow commute or a cold day could, although unlikely to be expressed along with actual software issues, could still skew any emotions attached with an actual issue.

The next two discussion points were very similar in that the usage of this idea is incredibly important. Context behind statements is lost, such as previously mentioned external factors affecting emotions as well as multi sentence sentiments that provide context may not be captured by an LLM. All of these concerns is where I provide my viewpoint, which people seemed to agree with in that this is a great way to get an overview of what could be affecting the company but is not necessarily the case and as such a sanity check would be warranted. This is where I say this emotion extraction would be something that would lead to further discussions whether it be in a mass email or interviews with employees or otherwise.

A student had then brought up a possible further work in that OpenAI now has a feature to tune their preexisting model for specific usages. I was unsure how this worked exactly as I worried it crosses too far into the territory of training a model for a specific use case but I was told it should have been an easy a light process that does not require further training but only what it is already trained on.

Professor Godfrey caps off the discussion with a great macro viewpoint on the studies of LLMs in recent years. He compares LLMs to an invasive alien species that everyone must now rush to find how this could revolutionize any use case. As such, everyone would be revisiting old problems to see if LLMs can be applied even if it does not end up being a good idea and as such some papers can be very unsatisfying. Suffice to say, usages of LLMs are in an exploratory phase and although we have found some nice use cases for them, new ideas are constantly being explored and although most may be failures at the end of it all, everyone is seeking that gold nugget, or even, that silver bullet.