

GPT-2-Emoji: a pre-trained language model approach to emotion classification

Riccardo Giacomelli

Stanford CS224U

ricgiac@stanford.edu

Abstract

Emotions classification is an important part of human interaction and plays a key role in understanding human communication. From customer service oriented tools or health care focuses chat-bots, to on-line reviews and automatic filtering and composition, many applications can benefit from emotion enhanced models. We proposed an all-in-text approach based on new version of GPT-2 which is offering a wide and flexible spectrum of utilizations. The Classification problem is cast to an all in text completion task. The input text and the emotion label are concatenated in a fixed schema. Given a text prefix, GPT-2 will complete the sentence with the corresponding emotion. The compositional and training forward only paradigm of GPT-2 allows an intuitive and expandable interface similar to human way to compose sentences. Furthermore, after fine-tuning, compositional creativity of GPT-2 is preserved for generating new emotional states as *inspired* or *tired* not present in the original set. We focus in particular on GoEmotion dataset, the largest manually annotated dataset of 58k Reddit comments. The dataset presents a challenge in the high number of imbalanced emotion categories. Our model achieves state of the art and shows compositional creativity in emotion classification.

1 Authorship statement

The author Riccardo Giacomelli carried out all the model training, experiments and writing with no external help. The author declares no conflict of interest. The author thanks all NLU group. In particular Prof. Chris Potts for the idea on emotion prediction and useful discussion and Dora Demszky for useful suggestions. Omar Khattab suggested very interesting articles on language models. Author also thanks Google TFRC for TPU time.

2 Introduction

Emotions classification is an important part of human interaction and core to wide range of services e.g. [Jain and Kumar, 2020], [Oh et al., 2017]. Emotion classification is particularly challenging because of cultural differences leading to different perception and emotional expression [Labidi et al., 2019]. Several datasets have been published, e.g. [Bostan and Klinger, 2018], [CrowdFlower, 2016]. Existing datasets are either small or covers a limited taxonomy [Demszky et al., 2020]. Moreover classification in literature are coarse e.g. [Ekman, 1992]. We focus on GoEmotions containing 27 emotion categories. Different nuances of emotions represents a challenge, they results in large number of classes which can still be relevantly separated [Demszky et al., 2020]. Beyond prevailing emotions as approval, admirations, amusement, anger which represents majority of recorded emotional states, transitional states [Sudhof et al., 2014] are recorded in a small number, leading to unbalanced classes. The proposed approach describe application of language models (GPT-2) [Madotto et al., 2020] to emotion classification in an all-in-text approach. Compared to bidirectional models as BERT [Devlin et al., 2019], at every point in the text sequence, GPT-2 does not have access to future tokens on the right of current token. The unidirectionality (backward dependency only) mimics the human way to write/talk forward only [Liu et al., 2021]. The advantage is the model can generate completion for input text showing creativity skills as poetry and music composition [Zitelli, 2020], [Branwen, 2019], [Geerlings and Meroño-Peñuela, 2020]. In this specific task the emotion classification is generated as a completion of the input text. No external classification layers are added to the encoder only architecture and the classification is specified all in text. This approach is particularly targeted to the case of high number of emotions where building feature and specific models for different emotions would be challenging. The all-in-text approach allows to target the classification problem in a straightforward and extensible way. Furthermore the flexibility of

the text only approach leaves open possibility to extension for conditional composition. To target unbalanced dataset classical approaches as oversampling/undersampling and weighting scheme for samples to give more even contribution in the gradient update. More recent approaches [Jiang et al., 2019] specific for language models, originally targeted to avoid overfitting and achieving better generalization can be applied to unbalanced dataset as well. We explored the creativity of GPT-2 letting the model to predict emotion labels not constrained to the original set. GPT-2 created new labels added to the 27 already in the dataset as *inspired*, *tired*, *need* and *drunk*. The new labels results accurate.

3 Related works

Previous emotion classification approaches includes word vector representation [Pennington et al., 2014], BERT [Demszky et al., 2020] with additional classification layer attached to output layer of the [CLS] token and T5 model [Raffel et al., 2019]. Large language models (GPT-2/3) in-context learning allows a pre-trained model to classify emotions learning on few examples in a few and zero shots learning fashion [Madotto et al., 2020] without need to fine tuning. Conditional random fields [Sudhof et al., 2014] which allows to model transition emotional states and improve classification for transitional, minority classes. Few general approaches focus on the training method to improve transfer learning and fine tuning [Si et al., 2009], [Jiang et al., 2019]. Our approach is based on fine tuning of a new pre-trained version of GPT-2 (Neo) in an all in text approach. The model is trained as a standard language model, predicting next token only. Compared to BERT random masking training, allowing a more efficient data utilization. Differently from BERT this approach does not require additional external layers. We generate emotion labels as next token completion. We are adopting a GPT-2 encoder only architecture with 1.3B parameters compared to T5 which utilizes an encoder-decoder architecture.

4 Data

The GPT-neo model has been pre-trained on C4 dataset [Dodge et al., 2021] (800 GiB) for 1 epoch. C4 has been tokenized by fast GPT-2 tokenizer and stored in tfrecords file format. The main dataset for fine-tuning, GoEmotion [Demszky et al., 2020], contains 27 emotion labels and a neutral class. It is the largest manually annotated dataset of 58k English Reddit comments. Characteristic of GoEmotion is the unbalance number of classes in training set, 4130 for admiration, 2596 for approval and just

57 and 96 samples for pride and relief respectively. 3 new dataset have been generated oversampling minority classes to match the number of examples of admiration class (4130). The neutral class is not modified. The maximum oversampling rate is set to 3,6,10 times. Excluding the neutral class, the original maximum imbalance ratio between admiration and pride (72:1) is reduced respectively to 24:1,3.6:1,2.4:1.

5 Model

5.1 Language modelling for emotion prediction

For the emotion prediction task, we exploit the GPT-2 next token generative completion. The training examples are formatted as a concatenation of the text to be classified and the emotion label in a single sentence, e.g.:

This is a comment: *I think the project will go very well.* ***Emotion is optimism.***

A fixed sentence structure represented by the text in bold is completed with the post (first part) and the emotion (ending word). During training the model can learn the full sentence, whereas in prediction the model will complete the final word representing the emotion label. The model can learn to predict multiple sentence at once utilizing a sequential formatting [Donahue et al., 2020].

5.2 GPT-Neo

We release weights for a new version ¹ of GPT-2 model (1.3B parameters), based on GPT-Neo [Black et al., 2021], an implementation based on Mesh tensorflow [Shazeer et al., 2018]. The model has 16 heads, 24 layers, embedding dimension 2048. Activation function is GELU [Hendrycks and Gimpel, 2016]. The size has been chosen to still fit in 8GiB GPU memory. Compared to the original released version, our version has been pre-trained for longer timesteps (700k) on The Pile [Gao et al., 2020] and C4 [Dodge et al., 2021]. The model was trained on Google Cloud TPU [cloud] under the TFRC program [Center] on a v2.8-512 core pod. Main fine-tuning approaches are described below. The aim is to address both the high number of labels and the unbalanced characteristic of the dataset. Our baseline is BERT f1-macro score reported in [Demszky et al., 2020]. f1-macro is chosen to have an equal contribution across all the emotions. This is particularly important to test performance of the model across majority and minority classes.

⁰checkpoint pre-trained on The Pile

¹A demo and full parameters list is available at <https://github.com/riccardo247/GPT-2-emoji/blob/main/GPT-2-emoji.ipynb>

6 Methods

6.1 Original dataset

To best perform, model is trained by giving a single example per context window. The training set contains a total of 43410 samples. GPT-2 is trained to predict next token according to standard language model training. During training the model learns: (1) the fixed structure of the sentence, (2) the specific language of the posts, (3) to associate posts to emotion. Three epochs achieved the best f1-macro result reported in 2. Full precision, recall and f1 results are reported in table 5 for test set.

6.2 Unbalanced classes

The highly unbalance nature of the dataset advocates generating a more balanced dataset. As described in section 4, the model is trained on 3 more balanced version of the original dataset. Results are reported in table 1. Oversampling does not improve f1-macro score. Moreover f1-macro is lowering with increasing oversampling. The interpretation is oversampling is a balance between replicating samples in each category and overfitting. In case of 27 categories is difficult to find the right balance. Other resampling ratios are left for future research.

max oversampling	dataset size	f1 macro
3x	47183	0.462
6x	101900	0.447
10x	108410	0.439

Table 1: Balanced datasets

6.3 Full context training

Language context window of GPT-2 allows to train the model on a concatenation of multiple training sentences at once. Since the posts are short text, approximately 58 of them can be fit in a 2048 token window. The hypothesis is that the model can learn relationships among the classes and improve classification. E.g. the last examples have access to all previous examples. Moreover a single context window is engineered to contain always at least one example for each class. The downside is that GPT-2 can learn spurious relationships based on ordering of the examples or specific choice of samples. During training every context window contains up to 59 training sentences. The proportions of original number of samples or class is approximately maintained in each training sequence. Pride, grief, relief contributes with one sample each. Majority classes as approval, admiration and neutral contribute respectively with 3,5 and 13 samples. All together they sum up to 59 samples. Few sequences with longer posts contains less than 59 samples. 59 samples in

the sequence are generated with algorithm 1. N total examples are generated. Every emotion e contributes with n_e samples proportionally to the respective total number of samples in the training set. During prediction a prefix sequence of 28 concatenated samples (one for each emotion) is randomly sampled followed by the example to be classified. The approach is similar to few-shots learning

[Madotto et al., 2020]. For every example to be predicted, 5 random prefix sequence are chosen and the model generate 5 different predictions. Final predicted emotion is the majority vote of the 5 classifications. This multiple classification approach is introduced to mitigate the possible bias of the random sampling. Result is reported in table 1 labelled all-shots. The best f1-macro (0.395) is much lower than the vanilla training. Further investigation is needed.

Algorithm 1 Full context sampling

```
1: while  $n \leq N$  do
2:    $samples = []$ 
3:   while  $e \leq 28$  do
4:      $samples += [n_e \text{ samples for class } e]$ 
5:      $e += 1$ 
6:   end while
7:    $gen\ sample_n = shuffle(samples)$ 
8:    $n += 1$ 
9: end while
```

experiment	batch size	epochs	learning rate
original	128	3	2e-04
balanced	128	3	2e-04
all-shots	128	3	2e-04

Table 2: Experiments summary table

6.4 Creative classification

In a further experiment we train the model on the original dataset, one sentence per context window. Differently than in first experiment we do not include neutral examples. In few cases the model predicts new emotions not in the original set. E.g. *Also feeling very inspired....i will definitely try doing the same thing.* has been classified as inspired.

7 Results

In table 3 we report the results of training the model using original training set, more balanced dataset and a sequence of sample to fill the 2048 tokens window. The best performing training method is the utilization of original training set, one example per context window, without oversam-

pling. The model achieves a state of the art f1-macro of 0.467 on the test set. This value is higher to 0.46, the one reported in [Demszky et al., 2020]². Both the upsampling and the sequence of sample approaches do not improve the performance.

experiment	f1 macro dev	f1 macro test
original	0.473	0.467
balanced	0.462	-
all-shots	0.395	-

Table 3: Experiments summary table

8 Analysis

Varying the batch size as shown in 4 keeping fixed learning rate to 2e-04 plays a role in performance. Best f1-macro results are achieved with batch size 128. The full results of best model in 5 show the model have best f1 for amusement compared to BERT in [Demszky et al., 2020]. Desire, love, surprise and remorse (0.392, 0.70, 0.365, 0.562) are much lower than the ones reported for BERT (0.49, 0.78, 0.5, 0.66). The score in dev is better for both love and surprise. Surprisingly for grief the model scored 0.4 whereas BERT scored 0. Giving high number of classes, variation of results is expected. Overall the model is performing well showing it is able to recognize different types of motions in particular for minority classes. Poor results for love and surprise have to be investigated further. One possible reason for poor performance of all-shots approach is the model has to learn both the language and the labels as next token prediction. Giving labels higher weights could help during training.

Batch size	f1 macro
64	0.46
128	0.473
192	0.461

Table 4: Batch size effect

9 Conclusion

We introduce an all-in-text approach for emotion classification. Different from few-shots learning approach, the new weights for GPT-2 Neo are pre-trained on The Pile and C4 datasets and then fine tuned on the GoEmotion dataset. The classification task is cast to a sentence completion task. The input text and the corresponding emotion category are concatenated in a single text sentence on which the model is trained. The model can achieve state of the art performance on the GoEmotion

²Although within 1 standard deviation

emotion	precision	recall	F1
admiration	0.688	0.708	0.698
amusement	0.758	0.869	0.810
anger	0.500	0.434	0.465
annoyance	0.351	0.297	0.322
approval	0.539	0.301	0.386
caring	0.331	0.377	0.352
confusion	0.520	0.367	0.430
curiosity	0.445	0.644	0.526
desire	0.714	0.270	0.392
disappointment	0.341	0.236	0.279
disapproval	0.383	0.423	0.402
disgust	0.385	0.417	0.400
embarrassment	0.750	0.300	0.429
excitement	0.514	0.419	0.462
fear	0.701	0.635	0.667
gratitude	0.813	0.850	0.831
grief	0.500	0.333	0.400
joy	0.529	0.548	0.538
love	0.604	0.858	0.709
nervousness	0.375	0.188	0.250
neutral	0.596	0.717	0.651
optimism	0.636	0.408	0.497
pride	0.500	0.250	0.333
realization	0.733	0.101	0.177
relief	1.000	0.143	0.250
remorse	0.581	0.543	0.562
sadness	0.638	0.407	0.497
surprise	0.482	0.293	0.365
F1 macro			0.467

Table 5: Test set precision, recall and F1 macro values

dataset. Moreover we explored different methods to train the model for the task. Upsampling minority classes to generate a more balanced dataset did not improve performance. We explored using a prefix of a random samples during prediction which did not improve result. To test the compositional creativity of the model we train the model excluding the neutral category. The model was able to create new emotions labels as *inspired*, *tired*, *need*. The proposed approach is intuitive and flexible to be used in different applications. In conclusion the forward only generative approach of GPT-2 encoder only is more parameter efficient than T5 and revealed to be better performing than BERT. Retained compositional skills benefit a variety of applications.

10 Future research

As suggested in [Donahue et al., 2020] is possible to train the model to predict just the last word (emotion label) without including previous words in the likelihood. This approach is left for future

research. Approaches to mitigate the unbalanced categories issue are sample weight inverse proportional to belonging class. Another approach described in [Jiang et al., 2019] which can allow the model to better learn classification examples in minority classes. Using the full context window both in training and prediction revealed to perform much worse than single sample approach. Emotional text generation can be obtained switching roles between labels and text during training. It is interesting to check if the model can generate text conditioned to multiple emotions in a combination not present in the training data. Further pre-training the model on additional reddit posts can be useful to learn the language before fine-tuning.

11 Impact statement

The approach represents an intuitive, simple and extensible method to emotion classification, and can inspire new research into the creative composition conditioned on emotions. The pre-training of GPT-neo model took about 1 week. Long training can have impact on the environment. The benefits of the model are more understanding of human emotions in text. This is beneficial in many ways and can help on mental health as well. The risk is a wrong utilization of the model to classify emotions in an unethical way. The model has been tested on a particular dataset biased to a specific Reddit posts and should not be considered a generalization. This could cause damage to single individuals and to society in general.

References

- Black, S., Gao, L., Wang, P., Leahy, C., and Biderman, S. (2021). GPT-Neo: Large scale autoregressive language modeling with mesh-tensorflow.
- Bostan, L.-A.-M. and Klinger, R. (2018). An analysis of annotated corpora for emotion classification in text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2104–2119. Association for Computational Linguistics.
- Branwen, G. (2019). Gpt-2 folk music.
- Center, T. R. *TFRC*.
<https://www.tensorflow.org/tfrc>.
- cloud, G. *GC*. <https://cloud.google.com/>.
- CrowdFlower (2016). *CrowdFlower*.
<https://data.world/crowdfower/sentiment-analysis-in-text>.
- Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., and Ravi, S. (2020). Goemotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.
- Dodge, J., Sap, M., Marasovic, A., Agnew, W., Ilharco, G., Groeneveld, D., and Gardner, M. (2021). Documenting the english colossal clean crawled corpus. *arXiv preprint arXiv:2104.08758*.
- Donahue, C., Lee, M., and Liang, P. (2020). Enabling language models to fill in the blanks. *arXiv preprint arXiv:2005.05339*.
- Ekman, P. (1992). Are there basic emotions?
- Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., et al. (2020). The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.
- Geerlings, C. and Meroño-Peñuela, A. (2020). Interacting with gpt-2 to generate controlled and believable musical sequences in abc notation. In *Proceedings of the 1st Workshop on NLP for Music and Audio (NLP4MusA)*, pages 49–53.
- Hendrycks, D. and Gimpel, K. (2016). Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.
- Jain, V. K. and Kumar, S. (2020). Predictive analysis of emotions for improving customer services. In *Natural Language Processing: Concepts, Methodologies, Tools, and Applications*, pages 808–817. IGI Global.
- Jiang, H., He, P., Chen, W., Liu, X., Gao, J., and Zhao, T. (2019). Smart: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. *arXiv preprint arXiv:1911.03437*.
- Labidi, A., Ouamani, F., and Saoud, N. B. B. (2019). Emoculture: Towards an ontology for describing cultural differences when expressing, handling and regulating emotions. In *KEOD*, pages 298–304.
- Liu, X., Zheng, Y., Du, Z., Ding, M., Qian, Y., Yang, Z., and Tang, J. (2021). Gpt understands, too. *arXiv preprint arXiv:2103.10385*.
- Madotto, A., Liu, Z., Lin, Z., and Fung, P. (2020). Language models as few-shot learner for task-oriented dialogue systems. *arXiv preprint arXiv:2008.06239*.

- Oh, K.-J., Lee, D., Ko, B., and Choi, H.-J. (2017). A chatbot for psychiatric counseling in mental healthcare service based on emotional dialogue analysis and sentence generation. In *2017 18th IEEE international conference on mobile data management (MDM)*, pages 371–375. IEEE.
- Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.
- Shazeer, N., Cheng, Y., Parmar, N., Tran, D., Vaswani, A., Koanantakool, P., Hawkins, P., Lee, H., Hong, M., Young, C., et al. (2018). Mesh-tensorflow: Deep learning for supercomputers. *arXiv preprint arXiv:1811.02084*.
- Si, S., Tao, D., and Geng, B. (2009). Bregman divergence-based regularization for transfer subspace learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(7):929–942.
- Sudhof, M., Gómez Emilsson, A., Maas, A. L., and Potts, C. (2014). Sentiment expression conditioned by affective transitions and social forces. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1136–1145.
- Zitelli, J. (2020). 345m-gpt-2 after james wright: Can ai generate convincing contemporary poetry?