App2Check at EMit: Large Language Models for Multilabel Emotion Classification

Gioele Cageggi[1], Emanuele Di Rosa[2] and Asia Uboldi[3]

[1] *Data Scientist at App2Check srl, Via XX Settembre, 14 - 16121, Genoa, Italy*
[2] *Chief Technology Officer at App2Check srl, Via XX Settembre, 14 - 16121, Genoa, Italy*
[3] *Data Scientist at App2Check srl, Via XX Settembre, 14 - 16121, Genoa, Italy*

Abstract

In this paper we compare the performance of three state-of-the-art LLM-based approaches for multilabel emotion classification: fine-tuned multilingual T5 and two few shot prompting approaches: plain FLAN and ChatGPT. In our experimental analysis we show that FLAN T5 is the worst performer and our fine-tuned MT5 is the best performer in our dev set and, overall, is better than ChatGPT3.5 on the test set of the competition. Moreover, we show that MT5 and ChatGPT3.5 have complementary performance on different emotions and that A2C-best, our unsubmitted system that combines our best performer models for each emotion, has a macro F1 that is 0.02 greater than the winner of the competition in the out-of-domain benchmark. Finally, we suggest that a perspectivist approach is more suitable for evaluating systems on emotion detection.

Keywords

Emotion Detection, Large Language Model, ChatGPT, FLAN, mT5, Prompt Engineering

1. Introduction

Categorical Emotions Detection refers to the machine learning task of detecting the presence of specific emotions in a text. Detecting customers emotions, for example, is a useful task having many practical applications in industry, from customer experience analysis to customer churn prevention.

The categories of emotions used may vary. In this paper we consider the 8 main emotions of Plutchik's wheel [2] (anger, expectation, disgust, fear, joy, sadness, surprise, trust), plus the emotion "love," which is one of the dyads, according to the Emit 2023 competition [3], and Neutral, which is absence of emotions.

In this paper, we:

1. present three approaches for detecting emotions in a text, all based on large language models (LLM)
2. show that, on the dev set, FLAN T5 is the worst performer and our fine-tuned MT5 is the best performer
3. overall, between our models, MT5 is better than ChatGPT3.5 on the dev and test set of the competition
4. show that MT5 and ChatGPT3.5 show complementary performance on different emotions
5. present A2C-best, the unsubmitted system that combines our best performer model for each emotion. A2C-best shows a macro F1 that is 0.02 greater than the winner of the competition in the out-of-domain benchmark
6. perform error analysis on instances where all systems disagree and re-annotate them, also showing that we disagree with some labels in the golden standard
7. suggest that a "perspectivist approach" [4] is more suitable for evaluating emotions detection systems.

This paper is structured as follows: after the introduction, we describe the three approaches taken into account, then, we present and discuss the results on both dev and test set of the competition. Finally, we present our conclusions.

2. Approaches Adopted

In this paper, we study two different approaches to solve the Categorical Emotion Detection task, both Transformer-based:

- LLM Fine-tuning: Starting from a pre-trained LLM model, we use the competition dataset to fine-tune the model in order to solve the specific task
- Few-Shot Prompting: Using an Instruction Tuned LLM, prompts are designed to properly guide the model in defining its behavior for the task.

Briefly, the main differences of these two approaches are:

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT [1]

gioele.cageggi@app2check.com (G. Cageggi);
emanuele.dirosa@app2check.com (E. Di Rosa);
asia.uboldi@app2check.com (A. Uboldi)
https://www.linkedin.com/in/emanueledirosa/ (E. Di Rosa)

- While fine-tuned models require a larger labeled dataset for training, prompt-based models work even with a smaller few-shot dataset
- Fine-tuning requires high computational and resource capacity to complete the training. Few-shot prompting focuses on the refinement of prompts and instructions without changing the model parameters
- The carbon footprint of the two approaches is quite different. Fine-tuning an LLM can be computationally expensive and energy-intensive. The environmental impact generally tends to be more energy-demanding than prompt tuning, which is considered more eco-friendly because it avoids a full-scale fine-tuning process
- Fine-tuned models achieve better accuracy values when there is an abundance of labeled data, while prompt tuning can offer reasonable performance even with a limited amount of labeled data
- While fine-tuned models make LLMs specialized for a specific task, prompt tuning allows for a more flexible approach to solving different tasks with minimal changes to the prompt.

Moreover, as internal reference, we build a system called A2C-Baseline. It combines multiple ML models, such as Decision Trees [5] and KNN models [6], where we select for each emotion the best one from a pool of models. The input text is vectorized using the tf-idf methodology.

Finally, we define a voting system, A2C-Voting, that combines the prediction of each sentence from A2C-mT5-r1, A2C-GPT-r2 and A2C-Baseline. It chooses for each prediction the result with the largest agreement. The majority is always guaranteed, being based on a binary ranking of individual emotions (present/not present) and a voting system on three different predictions.

2.1. Fine-tuned LLM

Fine tuning LLMs has been proved to be an effective approach for text classification problems and in [7] we showed to be the winner approach in all tasks of the ABSITA competition. MT5 [8] is the LLM we decide to use here. It is a multilingual variant of T5 [9], a text-to-text model released by Google in 2021. T5 uses a transformer-based architecture and can be fine-tuned to return text labels for classification tasks. MT5 has been pre-trained on mC4, which is a version of Common Crawl's multilingual web crawl corpus containing 101 languages. This enables the exploitation of the potential of the T5 model on a task involving Italian text.

In this paper, in order to use this model, we use the Hugging Face API [10] wrapped by the Simple Transformers [11] library. From the available models, we choose the google/mt5-base version, which has 580 million parameters. We tried to apply google/mt5-xxl, but out of memory exception prevents us from using it in a Google Colaboratory cloud environment. More specifically, has been trained using an Nvidia A100 GPU with 40GB of memory. Training is performed for 20 epochs on 90% of the competition training dataset with a stratified split strategy. In this paper this model is referred to as A2C-mT5-r1.

2.2. Plain FLAN

FLAN-T5 [12] is one of two Few-Shot Prompting approaches that we experiment with in this paper.

It is a model based on T5[9], on which we perform instruction fine-tuning. This process entails training the model using an instruction set that describes how to perform over 1000 additional tasks. The instruction fine-tuning process involves providing the model with an instruction set and executing the tasks specified in the instructions.

In this paper, we use Hugging Face's transformers library to import the google/flan-t5-xl model and use it. Then, through prompt engineering techniques, we develop a prompt to associate an input text with one or more emotions. In the first iteration of the solution, we use a single prompt to identify and associate all possible emotions if present in the input text. However, the model is not supporting this compact approach. Thus, we modify the prompt to identify a single emotion at a time. We find better outputs with this last approach. Then we develop ten prompts, one per emotion.

The prompts start with *Detect if the text provided contains EmotionX as emotion. If the emotion is available in the input text, the value will be 1; 0 otherwise* , where *EmotionX* is the emotion to look for. Then two sentences follow, one of which contains the emotion and the other does not. In this paper this model is referred to as A2C-FlanT5.

2.3. ChatGPT

ChatGPT 3.5 is the second of the two Few-Shot Prompting approaches that we apply to experiment with in this paper. The version we use in this model is gpt-3.5-turbo-0301 [13]. The specifics of the model have not been publicly disclosed yet. It is a similar model to the previous GPT-3 model [13], trained on a set of text and code created before Q4 2021. It is then trained using a reinforcement learning method with rewards derived from human comparison.

In this paper, we use the OpenAI library [14] to process requests to the model. Unlike the approach chosen for FLAN T5, we develop a prompt to simultaneously identify all emotions for each text input. We prepared

Table 1
Example of sentences where the systems and Human classification disagree with the Golden Standard.

Text	A2C Team	Gold	A2C-mT5-r1	A2C-GPT-r2
Mi bastano 5", ma se le esibizioni di Pannofino e Facciolini totalizzano più di 20 punti possiamo annullare l'edizione... #taleequaleshow	Disgust	Anticipation	Trust	Anger
RT @user: Ultimo attacca i giornalisti in sala stampa: "Me l'avete tirata". Clima tesissimo. #Sanremo2019	Anger	Neutral	Trust	Anger
Perché lo AMano #IMedici #IMedici3	Love	Neutral	Love	Love

a prompt with six examples of text inputs, taken from the competition training dataset. All emotions have been mapped within the text examples. The output requested within the prompt is structured as a JSON with as many keys as emotions, with a value of 1 if a given emotion is present, 0 otherwise.

The prompt used is the following: *Determin the emotions in the text provided, which is delimited by <>. The available emotions are: Anger, Anticipation, Disgust, Fear, Joy, Love, Neutral, Sadness, Surprise, Trust. Provide the answer in JSON format, with the following keys: Anger, Anticipation, Disgust, Fear, Joy, Love, Neutral, Sadness, Surprise, Trust. If that emotion is present inside the input text, the value will be 1; 0 otherwise.* A series of examples then follow, in the format: *Text: <...>Answer: {"Anger":0, "Anticipation":0, "Disgust":0, "Fear":0, "Joy":0, "Love":0, "Neutral":0, "Sadness":0, "Surprise":1, "Trust":1}*

Note that this model allows to identify all emotions simultaneously, unlike FLAN T5, in which emotions have been identified one at a time. In this paper, this model is referred to as A2C-GPT-r2.

2.4. Description of our best approach: A2C-best

A2C-mT5-r1 and A2C-GPT-r2 show to be complementary in their ability to accurately detect emotions in the evaluation sets. Specifically, in the dev set, A2C-mT5-r1 outperforms A2C-GPT-r2, while the latter exhibits better performance on Anger, Disgust, Fear, and Sadness. Based on these findings, in the following, we show A2C-best, which combines the top-performing A2C models for each individual emotion.

We show in 3.1 and 3.2 the results of its application ranking on the test set of the competition as unsubmitted system, since we believe that its results are interesting for the research community.

3. Experimental Analysis

In this paper, we refer to two types of datasets: the development dataset and the competition test set. The development dataset is used to select the best A2C models to submit to the competition, while the competition test set consists of both in-domain and out-of-domain data. The dev set is split from the competition training set using the stratified technique [15], which ensures that the original proportions of labels is maintained in each subset. The training is made on the 80% of the training dataset; models are selected on the 10% of the dataset and tested on the remaining 10%. Once models to submit have been selected, we retrained them on the 100% of the training data. From here on we will refer to *Dev set* as the model selection set; *In-domain test set* and *Out-of-domain test set* refer, respectively, as the in-domain and out-of-domain competition datasets.

Tables 2 and 3 show the A2C models that participated in the competition applied on the Dev set, but also additional models developed post-deadline, highlighted in italics for a fair detection. Tables 6 and 7 include all models from both A2C and other competitors applied on the competition test set. All tables display the Macro F1 and F1 metrics for individual emotions across all models.

3.1. Results on Dev Set

In Table 2 and 3, we show the results of our model on the Dev set, where unsubmitted models are shown in italic. The worst performer is A2C-FlanT5 with an MF1 of 0.27: it shows the worst performance on the Neutral label, with an F1 score of 0. The best performer between the models we evaluated for the submission is A2C-mT5-r1, with an MF1 of 0.45, showcasing better performance on 6 out of 10 emotions when compared to the models that are not highlighted in italic. For the second run, we decide to select A2C-GPT-r2 instead of A2C-Baseline, since it performs in a complementary way compared to A2C-mT5-r1, and to pursue a more innovative approach. More specifically, it is clear that A2C-mT5-r1 and A2C-GPT-r2 exhibit complementary performance on different emotions: A2C-GPT-r2 excels in Anger, Disgust, and Sadness, while A2C-mT5-r1 performs better in Anticipation, Joy, Neutral, Surprise, and Trust. This complementary performance is almost entirely preserved in the competition test sets as well. Based on this observation, we synthesize a post-deadline system called A2C-best which selects the model with the best performance for each emotion.

3.2. Results on Competition Test Sets

In domain test set In Tables 4 and 5, we compare both competitors systems and all our models on the in-domain Test Set of the competition. When we look at the individual emotions, ExtremITA run 2 achieves almost always the highest scores, except for Joy, where ABCD run 1 is the best one, and Love, where A2C-GPT-r2 is the best performer.

We also include in the tables the results obtained by A2C-best, which ranks second after ExtremITA's solutions, with an MF1 score at a distance of 0.005 from its first run.

The complementarity observed in the dev set between A2C-mT5-r1 and A2C-GPT-r2 also holds true within this test set, except for the emotion of Fear. We include an upper bound benchmark, Best-All, to define the potential margin of improvement by combining all the competition models.

Out-of-domain test set In Tables 6 and 7, we show the results of our systems and the other participants on the out-of-domain Test Set of the competition. Observing individual emotions, A2C-GPT-r2 shows the best score on Anger, Disgust, and Fear, while A2C-Voting on Sadness.

We obtain A2C-best by incorporating the best results of our models into one system, that selects the best of our models for each emotion. A2C-best shows to be the top performer among the submitted runs, improving the winner by 0.02 of MF1.

Once again, complementarity on emotions is clear between A2C-mT5-r1 and A2C-GPT-r2, except for Love. As an upper bound, Best-All shows that the potential margin of improvement is more significant in the out-of-domain test set.

Table 2
Dev set - performance of A2C models on emotions Anger, Anticipation, Disgust, Fear, Joy

MF1	Model	Ang	Ant	Dis	Fea	Joy
0.529	***A2C-best***	**0.47**	**0.60**	**0.55**	**0.40**	**0.52**
0.501	*A2C-Voting*	**0.47**	**0.60**	0.48	**0.40**	**0.52**
0.494	*A2C-r1r2*	0.41	0.56	**0.55**	0.29	0.45
0.447	A2C-mT5-r1	0.35	0.56	0.41	0.13	0.45
0.379	A2C-Baseline	0.34	0.52	0.37	0.31	0.29
0.360	A2C-GPT-r2	0.41	0.38	**0.55**	0.29	0.39
0.267	A2C-FlanT5	0.20	0.38	0.36	0.17	0.35

Table 3
Dev set - performance of A2C models on emotions Love, Neutral, Sadness, Surprise, Trust

MF1	Model	Lov	Neu	Sad	Sur	Tru
0.529	***A2C-best***	**0.55**	**0.55**	**0.54**	**0.48**	**0.62**
0.501	*A2C-Voting*	**0.55**	0.50	0.51	0.36	**0.62**
0.494	*A2C-r1r2*	0.51	0.55	**0.54**	**0.48**	0.59
0.447	A2C-mT5-r1	0.51	**0.55**	0.42	**0.48**	0.59
0.379	A2C-Baseline	0.34	0.36	0.38	0.32	0.57
0.360	A2C-GPT-r2	0.51	0.16	**0.54**	0.12	0.25
0.267	A2C-FlanT5	0.42	0.00	0.36	0.28	0.15

Table 4
In-domain Test set - performance of all systems on emotions Anger, Anticipation, Disgust, Fear, Joy

MF1	Model	Ang	Ant	Dis	Fea	Joy
0.608	*Best-All*	**0.52**	**0.64**	**0.63**	**0.58**	**0.64**
0.603	**extremITA2**	**0.52**	**0.64**	**0.63**	**0.58**	0.62
0.509	extremITA1	0.48	0.56	0.57	0.14	0.59
0.504	*A2C-best*	0.44	0.49	0.58	0.40	0.61
0.499	ABCD1	0.47	0.59	0.55	0.00	**0.64**
0.492	*A2C-r1r2*	0.40	0.41	0.58	0.40	0.61
0.484	E.Hunters1	0.46	0.52	0.58	0.24	0.46
0.452	A2C-mT5-r1	0.35	0.41	0.39	0.40	0.61
0.445	*A2C-Voting*	0.44	0.49	0.44	0.24	0.55
0.374	A2C-GPT-r2	0.40	0.38	0.58	0.26	0.36
0.329	A2C-FlanT5	0.25	0.34	0.38	0.33	0.34
0.299	A2C-Baseline	0.28	0.40	0.30	0.10	0.12

3.3. Error analysis

In order to improve our systems performance, we randomly selected instances in which all systems disagree to analyze the most difficult cases. However, during our error analysis, we noticed that many times we did not agree on the samples annotation. In table 1, we show just 3 samples (out of many) in which we disagree with the golden standard (two different people plus a referee). The goal is to highlight whether disagreement between the systems is due to just systems that cannot correctly meet the ground truth or if such instances may be interpreted in multiple ways and thus requiring multiple, equally correct, labeling. As we can see in table 1, there are differences between the Golden Standard (Gold column) and our classification (A2C team column). The research community is working towards the direction of perspectivist approaches (see [16] and [17]) in which, well-known issues of having just one single ground truth are taken into account especially in Natural Language

Table 5
In-domain Test set - performance of all systems on emotions Love, Neutral, Sadness, Surprise, Trust

MF1	Model	Lov	Neu	Sad	Sur	Tru
0.608	*Best-All*	**0.55**	**0.70**	**0.63**	**0.51**	**0.69**
0.603	**extremITA2**	0.52	**0.70**	**0.63**	**0.51**	**0.69**
0.509	extremITA1	0.45	0.66	0.52	0.42	**0.69**
0.504	*A2C-best*	**0.55**	0.53	0.52	0.33	0.58
0.499	ABCD1	0.46	0.65	0.60	0.38	0.65
0.492	*A2C-r1r2*	**0.55**	0.53	0.52	0.33	0.57
0.484	E.Hunters1	0.50	0.43	0.55	0.46	0.63
0.452	A2C-mT5-r1	0.49	0.53	0.43	0.33	0.57
0.445	*A2C-Voting*	0.54	0.51	0.42	0.24	0.58
0.374	A2C-GPT-r2	**0.55**	0.35	0.52	0.13	0.21
0.329	A2C-FlanT5	0.38	0.48	0.34	0.31	0.15
0.299	A2C-Baseline	0.34	0.40	0.20	0.30	0.55

Table 6
Out-of-domain Test set - performance of all systems on emotions Anger, Anticipation, Disgust, Fear, Joy

MF1	Model	Ang	Ant	Dis	Fea	Joy
0.564	*Best-All*	**0.64**	**0.60**	**0.68**	**0.18**	**0.44**
0.518	***A2C-best***	**0.64**	**0.60**	**0.68**	**0.18**	0.42
0.498	extremITA2	0.41	0.49	0.67	0.00	**0.44**
0.484	*A2C-r1r2*	**0.64**	0.43	**0.68**	**0.18**	0.42
0.449	extremITA1	0.50	0.37	0.62	0.00	0.32
0.438	*A2C-Voting*	0.39	**0.60**	0.65	0.00	0.25
0.402	A2C-mT5-r1	0.27	0.43	0.47	0.00	0.42
0.373	A2C-GPT-r2	**0.64**	0.33	**0.68**	**0.18**	0.25
0.303	A2C-Baseline	0.23	0.45	0.46	0.00	0.14
0.295	A2C-FlanT5	0.51	0.22	0.59	0.00	0.26

Table 7
Out-of-domain Test set - performance of all systems on emotions Love, Neutral, Sadness, Surprise, Trust

MF1	Model	Lov	Neu	Sad	Sur	Tru
0.564	*Best-All*	**0.76**	**0.64**	**0.44**	**0.41**	**0.86**
0.518	***A2C-best***	0.71	0.34	**0.44**	0.37	0.81
0.498	extremITA2	**0.76**	**0.64**	0.30	**0.41**	**0.86**
0.484	*A2C-r1r2*	0.65	0.34	0.32	0.37	0.81
0.449	extremITA1	0.73	0.56	0.20	0.34	0.85
0.438	*A2C-Voting*	0.71	0.31	**0.44**	0.24	0.79
0.402	A2C-mT5-r1	0.65	0.34	0.27	0.37	0.81
0.373	A2C-GPT-r2	0.64	0.26	0.32	0.14	0.30
0.303	A2C-Baseline	0.24	0.14	0.40	0.22	0.77
0.295	A2C-FlanT5	0.39	0.27	0.27	0.19	0.25

Processing (NLP), and propose multiple equally correct labeling samples. In our opinion, categorical emotions detection is one relevant example of NLP in which is very difficult to agree on just one golden standard.

4. Conclusion

In this paper we presented the systems runs we submitted at Emit 2023 competition for emotion detection in text, and also our post deadline system called A2C-best. In particular, we presented the performance of three different LLM-based approaches, such as fine-tuned multilingual T5, and two few shot prompting techniques, A2C-GPT-r2 and FLAN T5. Our A2C-best model shows significant improvement to our official run and comparable performance with the first ranker of the competition in the out-of domain run. A2C-best scores 0.099 below the winner in the in-domain run. Finally, after relabeling difficult instances where all systems and humans disagree, we suggested that a perspectivist approach is more suitable for evaluating systems on emotion detection.

References

[1] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[2] K. K. Imbir, Psychoevolutionary Theory of Emotion (Plutchik), Springer International Publishing, Cham, 2017, pp. 1–9. URL: https://doi.org/10.1007/978-3-319-28099-8_547-1. doi:10.1007/978-3-319-28099-8_547-1.

[3] O. Araque, S. Frenda, R. Sprugnoli, D. Nozza, V. Patti, EMit at EVALITA 2023: Overview of the Categorical Emotion Detection in Italian Social Media Task, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[4] V. Basile, F. Cabitza, A. Campagner, M. Fell, Toward a perspectivist turn in ground truthing for predictive computing, CoRR abs/2109.04270 (2021). URL: https://arxiv.org/abs/2109.04270. arXiv:2109.04270.

[5] L. Rokach, O. Maimon, Top-down induction of decision trees classifiers - a survey, IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 35 (2005) 476–487. doi:10.1109/TSMCC.2004.843247.

[6] C. D. M. et al., Introduction to information retrieval, Cambridge University Press, 2008. URL: https://nlp.stanford.edu/IR-book/pdf/irbookprint.pdf. doi:10.1017/CBO9780511809071.

[7] E. D. Rosa, A. Durante, App2check @ ate_absita

2020: Aspect term extraction and aspect-based sentiment analysis, in: V. Basile, D. Croce, M. D. Maro, L. C. Passaro (Eds.), Proceedings of (EVALITA 2020), volume 2765 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020. URL: https://ceur-ws.org/Vol-2765/paper122.pdf.

[8] L. X. et al., mt5: A massively multilingual pre-trained text-to-text transformer, in: K. T. et al. (Ed.), Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021, Associa tion for Computational Linguistics, 2021, pp. 483–498. URL: https://doi.org/10.18653/v1/2021.naacl-main.41. doi:10.18653/v1/2021.naacl-main.41.

[9] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, 2020. arXiv:1910.10683.

[10] Hugging Face website, 2023. URL: https://huggingface.co/.

[11] T. C. Rajapakse, Simple transformers, https://github.com/ThilinaRajapakse/simpletransformers, 2019.

[12] H. W. C. et al., Scaling instruction-finetuned language models, 2022. arXiv:2210.11416.

[13] L. O. et al., Training language models to follow instructions with human feedback, 2022. arXiv:2203.02155.

[14] Openai website, 2023. URL: https://openai.com/.

[15] K. Sechidis, G. Tsoumakas, I. P. Vlahavas, On the stratification of multi-label data, in: ECML/PKDD, 2011.

[16] F. Cabitza, , A. Campagner, V. Basile, Toward a perspectivist turn in ground truthing for predictive computing, Washington DC, USA, 2023.

[17] The perspectivist data manifesto, 2023. URL: https://pdai.info/.