This version of the contribution has been accepted for publication, after peer review but is not the Version of Record and does not reflect post-acceptance improvements, or any corrections. The Version of Record is available online at: https://doi.org/10.1007/978-3-031-60215-3_3. Use of this Accepted Version is subject to the publisher's Accepted Manuscript terms of use https://www.springernature.com/gp/open-research/policies/accepted-manuscript-terms.

Emotional Evaluation of Open-Ended Responses with Transformer Models

Alejandro Pajón-Sanmartín[1], Francisco de Arriba-Pérez[1], Silvia García-Méndez[1], Juan C. Burguillo[1], Fátima Leal[2], and Benedita Malheiro[3,4]

[1] Information Technologies Group, atlanTTic, University of Vigo, Campus Universitario de Vigo, Lagoas-Marcosende, 36310, Vigo, Spain
[2] REMIT, Universidade Portucalense, 4200-072 Porto, Portugal
[3] ISEP, Polytechnic of Porto, Rua Dr. António Bernardino de Almeida, 431, 4249-015 Porto, Portugal
[4] INESC TEC, Campus da Faculdade de Engenharia da Universidade do Porto, 4200-465 Porto, Portugal
{apajon,farriba,sgarcia}@gti.uvigo.es, J.C.Burguillo@uvigo.es, fatimal@upt.pt, mbm@isep.ipp.pt

Abstract. This work applies Natural Language Processing (NLP) techniques, specifically transformer models, for the emotional evaluation of open-ended responses. Today's powerful advances in transformer architecture, such as ChatGPT, make it possible to capture complex emotional patterns in language. The proposed transformer-based system identifies the emotional features of various texts. The research employs an innovative approach, using prompt engineering and existing context, to enhance the emotional expressiveness of the model. It also investigates spaCy's capabilities for linguistic analysis and the synergy between transformer models and this technology. The results show a significant improvement in emotional detection compared to traditional methods and tools, highlighting the potential of transformer models in this domain. The method can be implemented in various areas, such as emotional research or mental health monitoring, creating a much richer and complete user profile.

Keywords: Emotional analysis, GPT-3.5, spaCy, Transformers, Prompt engineering, Language-based Emotion Recognition

1 Introduction

Natural Language Processing (NLP) models have gained significant importance in recent years, primarily due to advances in their ability to analyse and understand human language in an automated manner [6]. These improvements have been propelled and the technology popularised by companies such as OpenAI[5] and Google[6].

[5] Available at https://openai.com, reviewed in January 2024.
[6] Available at https://ai.google, reviewed in January 2024.

This research uses NLP techniques for emotional evaluation of open-ended responses. To achieve this goal, transformer models with power to understand the human psyche are employed [7], becoming valuable tools in a wide range of fields such as psychology [8], education [12], stock market [17], or marketing [4].

The proposed method explores the ChatGPT model (GPT-3.5) [2] together with prompt engineering (as discussed in Section 4.1) to control response generation. These techniques are still experimental, as there are no rules specifying an exact input/output relationship, only recommendations to consider when using a model. Tests are conducted using spaCy for polarity detection, concluding with a comparison between both models. The aim is to provide an effective and straightforward solution for emotion detection to be deployed in any domain with minimal modification, something that does not currently exist.

This paper is structured as follows. Section 2 delves into the state of the art in natural language processing, Section 3 focuses on the objectives, and Section 4 details the methodology for analysing conversations using ChatGPT and spaCy. Section 5 briefly explores the applications of this approach for analysing extensive texts, highlighting the differences between the models. Finally, Section 6 presents the conclusions and outlines future steps.

2 Related work

NLP has grown significantly in recent years, primarily due to its flexibility to adapt to a wide range of applications and to generate complex responses with minimal instructions. This research explores the ability of NLP models to perform emotional evaluation, aiming to comprehend and analyse the emotions expressed in a text. The most prominent techniques are as follows.

Linguistic feature extraction, used by initiatives such as the Semantic Orientation Calculator [14], assigns polarities to different words, creating a dictionary, and applies several algorithms to calculate emotional scores for each entry, resulting in a final classification.

Supervised machine learning employs classification algorithms, such as Naive Bayes [19], Support Vector Machines [20], or Neural Networks [18], to train models with large labelled data sets to classify texts into emotional categories.

The most advanced and sophisticated solutions are based on transformers, which capture representations of words and contexts in a general and flexible way, adding significant value. Prominent models in emotional detection include Bidirectional Encoder Representation Transformers (BERT) [1] and Generative Pre-trained Transformers (GPT) [9], among others. Most of these solutions are proprietary, such as Anthropic[7], the basis for the enterprise conversational assistant Claude[8], or Inflection [9], a model used to create a personal intelligence

[7] Available at https://www.anthropic.com/, reviewed in January 2024.
[8] Available at https://www.anthropic.com/index/introducing-claude, reviewed in January 2024.
[9] Available at https://inflection.ai/about, reviewed in January 2024.

assistant. There are also open source approaches, such as Vicuna[10], a modified version of the Large Language Model Meta AI (LLaMA) [15] from Meta.

Given the high flexibility of transformer-based models, they have been applied in various domains, including virtual assistants (BingChat or ChatGPT), machine translation [13], automatic text summarising [5], semantic search [10], and emotional analysis [11]. However, most of these solutions are code-oriented, which significantly restricts their reuse by researchers or professionals. The primary advantages of the current proposal over existing methods lie in its simplicity and seamless adaptability to diverse environments. Furthermore, it harnesses advanced and diverse models, ensuring a high level of precision.

3 Objectives

The main objective of this project is to develop a robust system capable of analysing emotions and polarities in any text or human interaction with the greatest possible accuracy. Specifically, it aims to:

- **Detect emotions and polarities with GPT-3.5:** The GPT-3.5 NLP model will be used to: (*i*) detect and classify emotions and polarities; and (*ii*) identify the most prominent *topics* in each interaction for user profiling.
- **Detect polarities with spaCy:** The spaCy NLP library will be used to detect and analyse existing polarities.
- **Compare spaCy and GPT-3.5 results:** Comparisons will be made regarding performance, *i.e.* the ability to correctly capture text polarity.

4 Proposed method

The designed method explores prompt engineering with the help of GPT-3.5 and spaCy in two contexts: interactive conversations and extensive texts. The considered emotions – Joy, Anger, Aversion, Sadness, Surprise, Fear, and Neutral – are based on the primary emotions model of Ekman and Cordaro [16], whereas polarity includes the Negative, Positive, and Neutral labels.

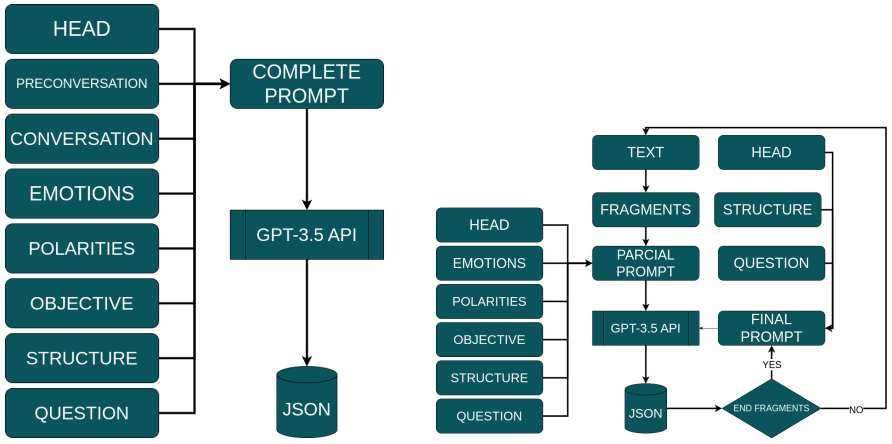
4.1 Prompt engineering

Prompt engineering, a technique currently under development [3], is used to control the results of generative artificial Intelligence (AI) models. A prompt is then a natural language textual description of the instructions or keywords given to guide a generative AI model. As a central tool in natural language processing, it can be used in a wide variety of applications.

[10] Available at https://lmsys.org/blog/2023-03-30-vicuna/, reviewed in January 2024.

4.2 Use of GPT-3.5

In a memory based approach, conversations are used as input data. The entire conversation is submitted to the model, indicating in the prompt the input to analyse in each iteration. This decision to provide the maximum possible context allows the model to obtain deeper results since past events are important [21]. To create a modular system, the iterative prompt creation process divides the text into eight fragments. Since the goal is to successfully detect emotions, polarities and keywords, the model is explicitly provided with the objective and parameters of the analysis (Figure 1a).



(a) Conversations (b) Texts

Fig. 1. Diagram of the analysis process with GPT-3.5

The analysis of a conversation, according to Figure 1a, includes the following steps: (*i*) assembly of all the fragments of the prompt; (*ii*) addition of the complete conversation; and (*iii*) specification of the conversation interaction to be analysed by the model (question). Then, there are other prompt fragments used to control the model and obtain the desired output, *e.g.*, emotions, polarities, objectivity, and structure. This process is repeated for each interaction[11] belonging to the conversation. The data are submitted via OpenAI's Application Programming Interface (API). Requests are made recursively until a satisfactory response is obtained. The results obtained by the model are processed and stored in a JSON[12] file, containing the polarity, emotion, and topics of each interaction.

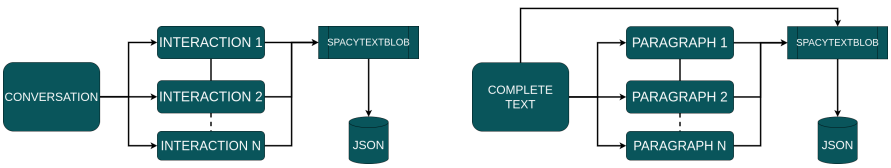
The analysis of extensive texts, *e.g.*, from interviews or transcriptions, is summarised in Figure 1b. In many cases, due to the limitation of 4096 *tokens*[13] per

[11] A conversation comprises one or more interspersed interactions between speakers.
[12] Available at https://www.json.org/json-en.html, reviewed in January 2024.
[13] Word used to designate a set of input characters submitted to the model.

model of GPT-3.5, it is impossible to segment texts based on paragraphs. Therefore, the text is dynamically divided into fewer fragments than this maximum. The prompt is also adapted, while maintaining the same modular philosophy. The process of analysing each fragment is identical to that of the conversations, sending each fragment sequentially. However, due to the *tokens* limitation, it is impossible to add the full text along with the fragment at this stage. In the last step, the JSON file holding the analysis of all the fragments is sent with a specific prompt that allows the data to be extracted from the full text. As a result, the final JSON file holds the complete analysis of the text. A dynamic adjustment of the request size has been performed. In the first call, the maximum size of 4096 *tokens* is maintained; however, after five requests, it automatically adjusts to the average between the total length of the prompt and the response.

4.3 Use of spaCy

The polarity analysis was performed with spaCy[14], a Python library for NLP. Specifically, it employs a light and popular polarity detection pipeline for English texts – the `spaCyTextBlob`[15] – based on the `TextBlob` library. The adopted pipeline loads the medium-sized English model trained on written web text. Therefore, all entries must be translated into English.



(a) Conversations (b) Texts

Fig. 2. Diagram of the analysis process with spaCy

The mode of operation is very similar to the one adopted with GPT-3.5. First, spaCy processes the interactions of a conversation. In this case, the process is simplified by not having to create the prompt, since the only input is the text, as can be seen in Figure 2a. Next, the model generates an output between -1 and 1, which is interpreted as positive, negative, or neutral if it is 0. Finally, the output is stored in a JSON file. In the case of the longer and more complex TED talk texts, they were translated for greater precision beforehand using the `textBlob`[16] translation API. Then, spaCy processes the entire text and the paragraphs independently since there has no entry size limitation.

[14] Available at https://spacy.io, reviewed in January 2024.
[15] Available at https://spacy.io/universe/project/spacy-textblob reviewed in January 2024.
[16] Available at https://textblob.readthedocs.io/en/dev/, reviewed in January 2024.

5 Experiments and results

The results were evaluated by calculating the weighted average of the *Precision*, *Accuracy*, *Recall* and *F1-score* metrics. To facilitate the interpretation of the results, the respective confusion matrices are also presented.

Since spaCy only detects polarity, the emotions and polarity analysis were performed independently with the Conversations[17], TED talks[18], and the Short phrases[19] data sets. Table 1 details the contents of these data sets. Prior any experiments, the Conversations and TED talks data sets were manually labelled by the authors, who have an NLP background.

Table 1. Contents of the data sets

Data set	Labels	Number of Entries	Words per Entry		
			Average	Minimum	Maximum
Conversations	No	1494	14	1	153
TED talks	No	2801	17	1	199
Short phrases	Yes[1]	14 448	9	3	34

[1] Includes only Positive and Negative polarity labels.

5.1 Emotion analysis

The emotion analysis was performed with the three data sets, using GPT-3.5[20]. The Conversations data set comprises an average of 14 words per interaction. As can be seen in Figure 3a, the results are positive since no attempt has been made to identify a contrary emotion, achieving an average *F1-score* of 71 %. The metrics improve with the simplest emotions, such as Joy or Fear, and fall moderately in more complex ones, such as Surprise.

The Spanish TED talks transcriptions contain an average of 17 words per paragraph. The results (Figure 3b) are more stable than the previous ones. Since the fragments are longer, on average 1735 words per talk, they provide more context, improving the emotional analysis to an average *F1-score* of 84 %.

Finally, the method was applied to the short phrases data set with an average of 9 words per sentence. The complexity of this analysis was higher since these are

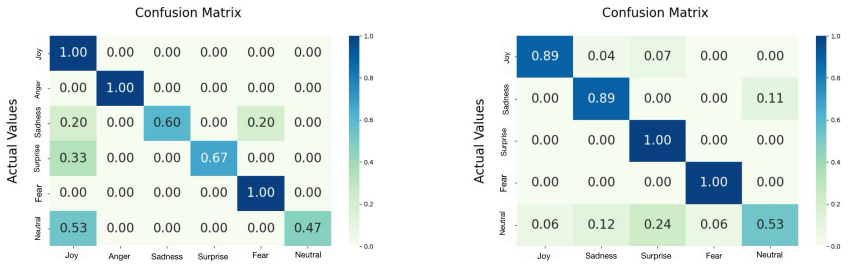
[17] Available at https://www.kaggle.com/datasets/projjal1/human-conversation-training-data, reviewed in January 2024.

[18] Available at https://www.kaggle.com/datasets/miguelcorraljr/ted-ultimate-dataset, reviewed in January 2024.

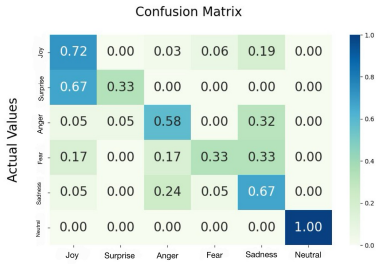
[19] Available at https://huggingface.co/datasets/hita/social-behavior-emotions, reviewed in January 2024.

[20] The model hyperparameters were set to `temperature=0.0`, `top_p=1.0` (default value), `frequency_penalty=0.0` (default value), `presence_penalty=0.0` (default value) and `stop_sequence=None` (default value).

phrases that have an inherent feeling, which results in a very complex labelling. Figure 3c shows that the results are positive and the wrong classifications are easily explainable. In complex emotions such as Surprise, confusion occurs with Joy, something that is not strange given the close relationship of these feelings. These disturbances cause performance to drop to an average *F1-score* of 62 %.



(a) Conversations (b) TED talks



(c) Short phrases

Fig. 3. Emotional analysis confusion matrices

These results, summarised in Table 2, show an *F1-score* performance above 70 % with two of the data sets. The majority of the wrongly labelled emotions correspond to related feelings, *i.e.* few were classified as antonyms.

Table 2. Emotion analysis with GPT-3.5

Data set	***Precision***	***Recall***	***F1-score***
Conversations	0.85	0.72	0.71
TED talks	0.86	0.85	0.84
Short phrases	0.62	0.61	0.62

5.2 Polarity analysis

Polarities have also been identified on all data sets with GPT-3.5 and spaCy. In the case of conversations, both models present good results, with GPT-3.5 obtaining an average *F1-score* value of 78 % and spaCy 66 %. The performance of spaCy with the TED talks drops to an average *F1-score* of 40 %, while that of GPT-3.5 increases to 88 % (Figure 4). The larger and more complex fragments affect negatively spaCy, because ambiguity typically increases as the sentence size increases, and favour GPT-3.5, due to richer context.

Fig. 4. Polarity analysis results (*F1-score*) with GPT-3.5 (green) and spaCy (rose)

The Short phrases results deepen the differences between both models. GPT-3.5 scores an average *F1-score* of 87 %, while spaCy scores 69 %. However, neutral polarities were ignored because they were not labelled in this data set.

These results indicate that both GPT-3.5 and spaCy can be used for polarity detection. However, there is a clear distinction in their recommended scope, depending on the length of the input. spaCy can be employed with smaller text fragments and GPT-3.5 with longer ones.

Table 3. Polarity analysis with GPT-3.5

Data set	*Precision*	*Recall*	*F1-score*
Conversations	0.89	0.77	0.78
TED talks	0.89	0.89	0.88
Short phrases	0.85	0.88	0.87

Table 4. Polarity analysis with spaCy

Data set	*Precision*	*Recall*	*F1-score*
Conversations	0.66	0.67	0.66
TED talks	0.49	0.39	0.40
Short phrases	0.69	0.69	0.69

6 Conclusions

This work explores the use of NLP techniques for emotional evaluation on open-ended responses. To this end, it uses transformer models together with prompt engineering to control the generation of responses. The emotion (GPT-3.5) and polarity results (GPT-3.5 and spaCy) were obtained with three data sets.

GPT-3.5 displayed better performance when compared with spaCy, specially with long texts. Transformer models produce a very precise analysis thanks to the identification of semantic relationships and the size of their databases. The use of prompt engineering was more complex than expected, since a large number of tests with different instructions had to be resorted to obtain an output that met the expected results. It was possible to obtain very different results with practically identical instructions, leading to errors in the subsequent processing. These problems were solved with fine tuning, such as limiting the number of *tokens* or increasing the value of the `top_p` hyperparameter used by GPT-3.5 to control the diversity of the generated text.

While the `spaCyTextBlob` pipeline is limited to polarity detection, the GPT-3.5 model performs a complete analysis of emotions, polarities and topics, which is essential for comprehensive applications. This proposal stands out for its advantages, particularly its adaptability and ease of use for non-technical users.

The future expansion of this emotion detection groundwork offers opportunities to create detailed psychological profiles and improve the ability to recognise self-destructive behaviour and identify early signs of neurological disorders through the integration of specialised algorithms. It is also necessary to further study the application of this method to other use cases, as each model has different limitations.

Acknowledgement

This work was partially supported by: (*i*) Xunta de Galicia grants ED481B-2021-118 and ED481B-2022-093, Spain; and (*ii*) Portuguese national funds through FCT – Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) – as part of project UIDB/50014/2020 (DOI: 10.54499/UIDP/50014/2020 | https://doi.org/10.54499/UIDP/50014/2020).

References

1. Al-Omari, H., Abdullah, M.A., Shaikh, S.: EmoDet2: Emotion detection in English textual dialogue using BERT and BILSTM models. In: 2020 11th International Conference on Information and Communication Systems (ICICS). pp. 226–232. IEEE (2020)
2. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. In: Advances in Neural Information Processing Systems. vol. 33, pp. 1877–1901. Curran Associates, Inc. (2020)
3. Dale, R.: GPT-3: What's it good for? Natural Language Engineering 27(1), 113–118 (2021)
4. Hartmann, J., Netzer, O.: Natural language processing in marketing. In: Artificial Intelligence in Marketing, vol. 20, pp. 191–215. Emerald Publishing Limited (2023)
5. Kai, W., Lingyu, Z.: Research on text summary generation based on bidirectional encoder representation from transformers. In: 2020 2nd International Conference on Information Technology and Computer Application (ITCA). pp. 317–321 (2020)

6. Khurana, D., Koli, A., Khatter, K., Singh, S.: Natural language processing: State of the art, current trends and challenges. Multimedia tools and applications 82(3), 3713–3744 (2023)
7. Liu, Y., Zhang, Y., Wang, Y., Hou, F., Yuan, J., Tian, J., Zhang, Y., Shi, Z., Fan, J., He, Z.: A survey of visual transformers. IEEE Transactions on Neural Networks and Learning Systems pp. 1–21 (2023)
8. Mann, P., Matsushima, E.H., Paes, A.: Detecting depression from social media data as a multiple-instance learning task. In: 2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII). pp. 1–8 (2022)
9. Mao, R., Liu, Q., He, K., Li, W., Cambria, E.: The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection. IEEE Transactions on Affective Computing 14(3), 1743–1753 (2023)
10. Masuda, K., Matsuzaki, T., Tsujii, J.: Semantic search based on the online integration of NLP techniques. Procedia - Social and Behavioral Sciences 27, 281–290 (2011), computational Linguistics and Related Fields
11. Peng, S., Cao, L., Zhou, Y., Ouyang, Z., Yang, A., Li, X., Jia, W., Yu, S.: A survey on deep learning for textual emotion analysis in social networks. Digital Communications and Networks 8(5), 745–762 (2022)
12. Shen, J.T., Yamashita, M., Prihar, E., Heffernan, N., Wu, X., Graff, B., Lee, D.: MathBERT: A pre-trained language model for general NLP tasks in mathematics education (2023)
13. Sun, Z., Wang, M., Li, L.: Multilingual translation via grafting pre-trained language models (2021)
14. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M.: Lexicon-Based Methods for Sentiment Analysis. Computational Linguistics 37(2), 267–307 (06 2011)
15. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: LLaMA: Open and efficient foundation language models (2023)
16. Tracy, J.L., Randles, D.: Four models of basic emotions: A review of Ekman and Cordaro, Izard, Levenson, and Panksepp and Watt. Emotion Review 3(4), 397–405 (2011)
17. Wang, C., Chen, Y., Zhang, S., Zhang, Q.: Stock market index prediction using deep transformer model. Expert Systems with Applications 208, 118128 (2022)
18. Wehrmann, J., Becker, W., Cagnini, H.E.L., Barros, R.C.: A character-based convolutional neural network for language-agnostic twitter sentiment analysis. In: 2017 International Joint Conference on Neural Networks (IJCNN). pp. 2384–2391 (2017)
19. Wongkar, M., Angdresey, A.: Sentiment analysis using naive bayes algorithm of the data crawler: Twitter. In: 2019 Fourth International Conference on Informatics and Computing (ICIC). pp. 1–5 (2019)
20. Zainuddin, N., Selamat, A.: Sentiment analysis using Support Vector Machine. In: 2014 International Conference on Computer, Communications, and Control Technology (I4CT). pp. 333–337 (2014)
21. Zhou, Y., Kang, X., Ren, F.: Prompt consistency for multi-label textual emotion detection. IEEE Transactions on Affective Computing pp. 1–10 (2023)