

An Experimental Comparison of Large Language Models for Emotion Recognition in Italian Tweets

Claudia Diamantini^{1,†}, Alex Mircoli^{1,*,†}, Domenico Potena^{1,†} and Simone Vagnoni^{1,†}

¹Department of Information Engineering, Università Politecnica delle Marche, Ancona, Italy

Abstract

In recent years, the advent of Large Language Models (LLMs), which are task-agnostic models trained on huge amounts of textual data, has given momentum to a wide variety of NLP applications, ranging from chatbots to sentiment classifiers. Currently, many LLMs are publicly available, each with different features and performance, and the selection of the best LLM for a specific task may be challenging. In this work, we focus on the task of emotion recognition in Italian social media content and we present an experimental comparison among three of the most popular LLMs: Google Bidirectional Encoder Representations from Transformers (BERT), OpenAI Generative Pre-trained Transformer 3 (GPT-3) and GPT-3.5. Model specialization in emotion recognition has been achieved by using two different approaches, namely fine-tuning and prompt engineering with few-shot task transfer. The experimentation has been performed on TwIT, a corpus of about 3100 Italian tweets annotated with respect to six emotions. The results show that fine-tuning GPT-3 leads to the best performance on the considered dataset, achieving a remarkable $F_1=0.90$.

Keywords

emotion recognition, BERT, GPT-3, large language model, sentiment analysis, emotion recognition of tweets, emotion recognition in Italian, fine tuning, few-shot learning

1. Introduction

The advent of social networks has made available huge amounts of user-generated content, whose analysis could give valuable insights into people's feelings and opinions. For this reason, several techniques for the semantic analysis of natural language have been developed. Among others, *emotion recognition* algorithms have been proposed to analyze emotions expressed in texts. Such algorithms are usually developed through a supervised learning approach and hence, given the complex nature of textual data, require to be trained on enormous manually-annotated datasets, whose creation is costly and time-consuming. A first attempt to overcome this limitation is represented by the techniques for the automatic creation of annotated datasets by exploiting noisy indicators, such as emojis [1] or facial expressions [2]. Anyway, a more promising approach has been proposed in recent years thanks to the popularity gained by attention-based neural networks [3] [4] such as Transformers. These architectures mitigate

ITADATA2023: The 2nd Italian Conference on Big Data and Data Science, September 11–13, 2023, Naples, Italy


*Corresponding author.

†These authors contributed equally.

✉ c.diamantini@univpm.it (C. Diamantini); a.mircoli@univpm.it (A. Mircoli); d.potena@univpm.it (D. Potena); vagnonisimone96@gmail.com (S. Vagnoni)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

some of the challenges associated with Recurrent Neural Networks (RNNs) [5] and Long Short-term Memory (LSTM) networks [6] and led to the development of *Large Language Models* (LLMs). LLMs are massive Transformer-based neural networks that have been trained on huge amounts of data and offer unprecedented performance in a large variety of NLP activities. LLMs are usually general-purpose models and hence they are not specialized in a single task. In order to improve their performance on a complex task such as emotion recognition, a *fine-tuning* phase is usually needed. Such task consists of a small training on an emotionally annotated dataset, with the aim of correctly defining the eligible responses (i.e., the chosen classes) and showing the network some samples related to the chosen domain. Nevertheless, even if some works (e.g., [7]) have shown that fine-tuning an LLM on a small dataset is often sufficient to obtain good results in terms of classification accuracy, the customization of LLMs remains costly as it is usually performed through paid API calls. For this reason, in the present work we also investigate the effect of *prompt engineering* on LLMs. Prompt engineering is the name given to the process of determining the best sentence (also known as *prompt*) to ask LLMs in order to obtain the best possible response¹. In particular, LLMs demonstrated to give good results performing unknown tasks when a detailed description of the task in natural language and a few shot examples are included in the prompt. Such an approach has been investigated by Brown et al. [8], which found that model performance increased on several benchmarks, demonstrating that a task-agnostic LLM can be turned into a task-specific model through few-shot task transfer. Although LLMs have been widely tested on English corpora, limited experimentation has been conducted on other languages. For this reason, we focus on emotion recognition in Italian texts and, in particular, in Twitter data, since they are usually difficult to classify.

The contributions of the present work are two-fold:

- the experimental evaluation of three publicly-available LLMs on the emotion recognition task, in particular for Italian social media content. LLMs have been selected based on their popularity and the chosen models are: Google Bidirectional Encoder Representations from Transformers (BERT) [9], OpenAI Generative Pre-trained Transformers 3 (GPT-3) [8] and OpenAI GPT-3.5;
- the comparison between two different techniques, namely a traditional fine-tuning and a few-shot task transfer through prompt engineering.

The rest of the paper is structured as follows: the next section presents some relevant related work on emotion recognition. The description of the techniques used for fine-tuning and prompt engineering is proposed in Section 3, while Section 4 reports the results of the experimental evaluation of the models on a real-world dataset of Italian tweets. Finally, Section 5 draws conclusions and discusses future work.

2. Related work

In recent literature, much research effort has focused on the task of emotion recognition, applying it to different data types: audio [10], images [11], videos [12], and texts [13]. With respect to the latter, the majority of works on emotion recognition are limited to the English

¹<https://itnext.io/prompt-engineering-the-magical-world-of-large-language-models-dde7d8d043ee>

language [14] [15]. The main differences among such works are on the adopted emotional framework and the architecture of the proposed classifier. For what concerns the emotional frameworks, the most widespread are Ekman’s theory of six archetypal emotions [16] and Plutchik’s wheel of emotions [17]. With regard to the classification algorithm, researchers were mainly focused on word embeddings (e.g., Word2Vec [18]) until the advent of Transformer-based architectures. Such architectures have shown unprecedented performance and have rapidly replaced older approaches. In the last five years, a large number of Large Language Models based on Transformers have been released: among others, OpenAI GPT-3 [8] and Google BERT [19] have gained much popularity among researchers and practitioners. Despite the great availability of models for the English language (e.g., [20]), only a few resources for emotion recognition are available for other languages, including Italian. To the best of our knowledge, the two most recent LLM-based approaches for emotion recognition in Italian texts are [21] and [22]. In particular, in the first work the authors propose AIBERTO, which is a BERT-based LLM for the Italian language which has been created by fine-tuning BERT-Base on a large dataset of Italian tweets. Such work is similar to [22], where authors use emojis as noisy indicators to build an annotated dataset for fine-tuning BERT. For what concerns the experimental comparisons of LLMs for emotion recognition, in particular for the Italian language, to the best of our knowledge no relevant works have been published in the literature prior to the writing of this paper.

3. Methodology

The proposed methodology aims to evaluate the differences both between different LLMs and different task transfer techniques, in order to empirically determine the approach that may lead to the best performance in emotion recognition of Italian texts. The comparison between LLMs has been performed using the same task transfer technique, namely *fine-tuning*; the considered models are Google BERT and OpenAI GPT-3. For what concerns the comparison between different task transfer techniques, i.e. *fine-tuning* and *few-shot task transfer through prompt engineering*, the same version of the LLM could not be used due to API limitations in newer GPT versions. For this reason, the comparison has been made between the fine-tuned version of GPT-3 and the prompt-engineered version of GPT-3.5.

3.1. Fine-tuning

In this subsection we describe the approaches used to fine-tune the two considered LLMs. The selected BERT-based model for emotion recognition of Italian tweets, i.e. EmotionAIBERTO, is the result of fine tuning an Italian version of Google BERT, namely AIBERTO [21], which, in turn, has been created by fine-tuning BERT-Base on a dataset of about 200 million Italian tweets named TWITA[23]. Such a LLM has been developed without using the “next following sentence” technique, thus making it unsuitable for tasks like question answering but perfectly appropriate for emotion recognition. In order to fine-tune AIBERTO, we followed the approach presented in [22]. In particular, we added a final classification stage to AIBERTO and then we fine-tuned the entire network on the TwIT dataset. The entire architecture is depicted in Figure 1: the text is fed into AIBERTO, which generates a convenient sentence representation, and it is

then classified through a classification stage consisting of a fully connected layer and a softmax layer with 6 neurons (i.e., one for each considered emotion).

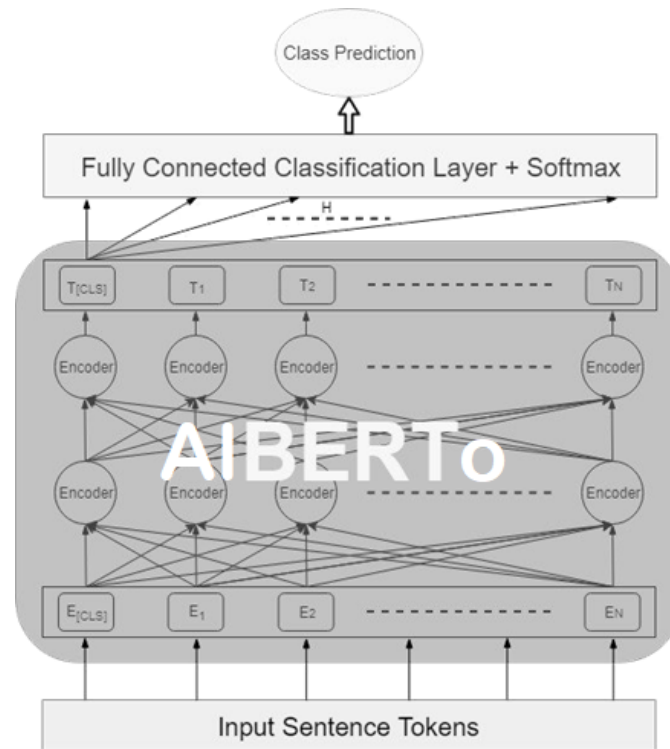


Figure 1: The architecture used for fine-tuning BERT [22].

For what concerns GPT-3, the fine-tuning has been performed through the OpenAI APIs². Such APIs require training data to be formatted in the JSON lines (JSONL) format, which is equivalent to JSON format but implemented using newline characters to separate JSON values. The training dataset must be converted so that each line is a *prompt-completion* pair where, in the case of emotion recognition, *prompt* corresponds to the sentence to be classified and the *completion* is the related emotion. The fine-tuned model is stored on the cloud of OpenAI and it is not downloadable: it can only be accessed through specific API calls in which it has to be explicitly selected as the current model. This aspect represents a potential limitation since the API calls, that are required for both training and inference, are paid and their cost depends on the number of analyzed tokens.

3.2. Prompt engineering

Few-shot task transfer through prompt engineering can be performed by defining an optimal initial prompt for the LLM, which may take into account both contextual information and training data. Regarding contextual information, it has been demonstrated that giving detailed

²<https://platform.openai.com/docs/guides/fine-tuning>

instructions about the task to be performed and the semantics of each considered class improves the model's ability to discriminate between emotions. For this reason, we started our prompts with an accurate description of the task and a description of each considered emotion. Subsequently, we add some *sentence-class* example pairs extracted from the dataset.

After several attempts (see 4.1), we found the following text to be the best initial prompt for GPT-3.5:

- ITA: *"Sei uno sociologo esperto nell'analisi delle emozioni espresse sui social network, che classifica le emozioni nel testo secondo questo schema: 1) felicità: Sentimenti di piacere, contentezza, soddisfazione, o anche attrazione e desiderio. Può includere risposte a complimenti o manifestazioni di affetto. 2) fiducia: Sentimenti di sicurezza, affidabilità o apprezzamento verso gli altri. Può comprendere la fiducia in se stessi, negli altri o nelle situazioni. Può anche includere sentimenti di rispetto o ammirazione per qualcuno o qualcosa. 3) tristezza: Sentimenti di dolore, malinconia o dispiacere. Può comprendere la delusione, il dispiacere per una perdita o un fallimento, o la sensazione di mancanza o vuoto. 4) rabbia: Sentimenti di frustrazione, irritazione o ira. Può includere reazioni a ingiustizie, insoddisfazioni o comportamenti negativi da parte degli altri. 5) paura: Sentimenti di preoccupazione, ansia o paura. Può comprendere la paura di eventi futuri, l'ansia per situazioni attuali o preoccupazioni in generale. 6) disgusto: Sentimenti di avversione, repulsione o disprezzo. include sentimenti verso comportamenti immorali, cibi o odori sgradevoli, o qualsiasi altra cosa che provoca una forte avversione. Nel contesto considera gli indizi lessicali, l'uso di simboli, di emoji, dell'ironia. Rispondi scegliendo soltanto una tra le seguenti emozioni: felicità, fiducia, tristezza, rabbia, paura, disgusto. [...]"*
- ENG: *"You are a sociologist expert in the analysis of emotions expressed on social networks, which classifies the emotions in the text according to this scheme: 1) happiness: Feelings of pleasure, contentment, satisfaction, or even attraction and desire. It may include responses to compliments or displays of affection. 2) trust: Feelings of security, trustworthiness or appreciation towards others. It can include trust in oneself, in others or in situations. It can also include feelings of respect or admiration for someone or something. 3) sadness: Feelings of pain, melancholy or sorrow. It can include disappointment, sorrow over a loss or failure, or feelings of lack or emptiness. 4) anger: Feelings of frustration, irritation or anger. It can include reactions to injustice, dissatisfaction or negative behavior from others. 5) fear: Feelings of worry, anxiety or fear. It may include fear of future events, anxiety about current situations, or worries in general. 6) disgust: Feelings of aversion, repulsion or contempt. includes feelings about immoral behavior, unpleasant foods or smells, or anything else that causes a strong dislike. In context, consider lexical clues, the use of symbols, emojis, irony. Answer by choosing only one of the following emotions: happiness, trust, sadness, anger, fear, disgust. [...]"*

For the sake of room, we have omitted the part of the text where the example sentences were provided.

4. Experiments

In this section, we discuss the result of an experimentation aimed at determining the best LLM and the best specialization technique for emotion recognition in Italian texts.

4.1. Experimental setup

The LLMs have been evaluated on the TwIT dataset [22], which is a dataset of 3108 Italian tweets labeled with regard to six emotions: happiness, trust, sadness, anger, fear, and disgust. The dataset is available at the following URL: <https://github.com/a-mircoli/twit>. The chosen emotions are consistent with those found in other datasets (e.g., MultiEmotionsIT [24]) and are considered basic, universal emotions, with the exception of *trust*, which has been added since it is quite common in social media texts. The class distribution is shown in Figure 2; it can be noticed that the dataset is quite balanced, with a maximum difference of 70 samples between the majority and the minority class.

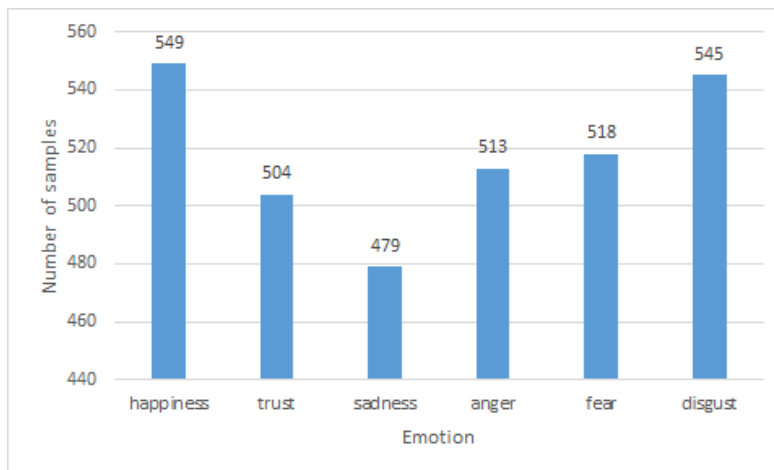


Figure 2: Class distribution of the TwIT dataset.

We tested the following LLMs:

- *BERT*: we considered EmotionAlBERTo, which is the fine-tuned version of AlBERTo, and we used the network hyperparameters shown in Table 1, since they provided the best results in a previous experimentation on the same dataset, as discussed in [22].
- *GPT-3*: we fine-tuned both the *davinci* model, which is the largest and costly GPT-3 model, and the *curie* model, which is smaller but faster than *davinci*. In the following, we only report the results of *davinci*, since it achieved slightly better results on the considered dataset.
- *GPT-3.5*: we performed few-shot task transfer through prompt engineering. In particular, we measured the performance of 20 different prompts, in which we varied the number of given examples and the context description, in order to find the best prompt. The results presented in the following subsection are related to the latter. We worked on the version

available in June 2023. It has to be noticed that this model receives frequent updates that may alter its performance and hence impact on the reproducibility of the obtained results.

Table 1

The optimal hyperparameters found for EmotionALBERTo through the hyperparameter tuning phase.

Hyperparameter	Value
learning_rate	2e-5
train_batch_size	512
eval_batch_size	512
max_seq_length	128
num_training_epochs	10

We evaluated the model by means of three metrics: precision, recall and F_1 score. Let x_{ij} be the number of data belonging to j -th class which have been classified as i -th class and let C be the number of classes. Precision and recall of i -th class are determined as follows:

$$precision_i = \frac{x_{ii}}{\sum_{j=1}^C x_{ij}} \quad (1)$$

$$recall_i = \frac{x_{ii}}{\sum_{j=1}^C x_{ji}} \quad (2)$$

F_1 score of i -th class is equal to:

$$F_{1i} = 2 \cdot \frac{precision_i \cdot recall_i}{precision_i + recall_i} \quad (3)$$

Therefore, the F_1 score achieved by a classification model is defined as the average of F_{1i} :

$$F_1 = \frac{1}{C} \sum_{i=1}^C F_{1i} \quad (4)$$

4.2. Results

The results of the experiments are shown in Table 2. It can be noticed that, even if the difference between the fine-tuned LLMs is quite small (4%), GPT-3 achieves the best performance, with a remarkable $F_1=0.90$. Conversely, the prompt-engineered version of GPT-3.5 achieves significantly lower results, settling on a F_1 score equals to 0.48. In particular, it can be seen that this model has great difficulties in classifying the *trust* emotion.

The confusion matrix for the best model is shown in Table 3. The confusion matrix highlights how the emotions sadness and anger are the most difficult to recognize, as they count respectively 14 and 13 misclassifications, while only 3 *trust* sentences are misclassified, leading to a very high precision (0.97) for the *trust* class. In general, the classification of the *happiness* and the *trust* classes seems to be easier for the model, as they have the highest values for precision and recall.

Table 2

Results of the experiments on the TwIT dataset. F_1 score and class-related precision and recall are reported for each model.

Model	F_1 score	Metric	Happiness	Trust	Sadness	Anger	Fear	Disgust	Avg
BERT	0.86	precision	0.95	0.96	0.82	0.79	0.86	0.76	0.86
		recall	0.96	0.97	0.89	0.71	0.80	0.81	0.86
GPT (fine-tuning)	0.90	precision	0.94	0.97	0.86	0.86	0.92	0.87	0.90
		recall	0.96	0.93	0.90	0.88	0.82	0.93	0.90
GPT (prompt eng.)	0.48	precision	0.45	0.25	0.53	0.45	0.73	0.61	0.51
		recall	0.91	0.11	0.57	0.71	0.15	0.37	0.47

Table 3

Confusion matrix for the best classifier: GPT-3 (fine-tuning)

	Act. Happiness	Act. Trust	Act. Sadness	Act. Anger	Act. Fear	Act. Disgust
Pred. Happiness	96	2	1	0	1	1
Pred. Trust	3	89	0	0	0	0
Pred. Sadness	2	0	78	3	8	1
Pred. Anger	0	3	2	110	2	6
Pred. Fear	0	1	6	3	93	2
Pred. Disgust	0	1	2	5	5	95
Recall	0.96	0.93	0.90	0.88	0.82	0.93
Precision	0.94	0.97	0.86	0.86	0.92	0.87

On the basis of the obtained results, it could be concluded that fine-tuning offers superior performance compared to prompt engineering, despite the fact that the latter was done on a newer and more performing version of GPT. In fact, it is possible to note that the fine-tuning causes the LLM to adapt very much to the concept of happiness, trust, etc. expressed in the training set, for which the classification is extremely better. However, it should be noted that the fine-tuning is done on TwIT, which is a dataset that was annotated based on the emojis present in the text, which were subsequently removed. This fact shouldn't be overlooked as emojis provide hints on how to classify ambiguous sentences such as "Un'altra cena di 3 ore? Gna faccio" (ENG: Another 3-hour dinner? I can't do it) which could be associated with various classes (e.g., sadness, fear, disgust) on the basis of the emojis added to the text. In this particular example, the emoji with the disgusted face (present when the sentence was collected and then removed after assigning the class to the sentence) adds a bias that GPT-3 is able to capture thanks to the fine-tuning done on the annotated dataset but GPT-3.5 does not, because the emoji was previously removed. Another example is represented by the sentence "Wtf? mi sento male" (ENG: Wtf? I feel bad), which changes meaning depending on whether it is coupled with a scared or smiling emoji.

5. Conclusion and future work

The goal of the work was the experimental comparison of three LLMs (i.e., Google BERT, OpenAI GPT-3 and GPT-3.5) on emotion recognition in Italian tweets. The LLMs were specialized on this

task following two different approaches, namely fine-tuning and prompt engineering with few-shot task transfer, in order to determine the best technique in terms of classification accuracy and training effort. The models were tested on TwIT, a corpus of 3100 Italian tweets labeled with respect to six emotions. The experimentation showed that fine-tuning GPT-3 leads to the best classification performance ($F_1=0.90$) and, in particular, is more capable of analyzing more complex and nuanced emotions like sadness and fear, suggesting that it is better able to capture semantic aspects of text. For what concerns the comparison between the two specialization approaches, fine-tuning produced significantly better results than prompt engineering ($F_1=0.48$). In future work, we plan to include in the experimentation some other popular LLMs, such as Google PaLM 2 and Meta LLaMA, in order to carry out a more complete comparison of the models on the market. We are also working on the creation of a larger manually-annotated dataset with the purpose of testing the models on a broader variety of topics. Finally, we are interested in delving into the emerging field of multilabel emotion recognition, where texts are labeled with multiple emotion classes, providing a more nuanced representation of emotions and considering the typical subtleties of human feelings.

References

- [1] J. Islam, R. E. Mercer, L. Xiao, Multi-channel convolutional neural network for twitter emotion and sentiment recognition, 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), 2019, pp. 1355–1365.
- [2] C. Diamantini, A. Mircoli, D. Potena, E. Storti, Automatic annotation of corpora for emotion recognition through facial expressions analysis, 2020, p. 5650 – 5657. URL: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85110414516&doi=10.1109%2fICPR48806.2021.9413311&partnerID=40&md5=25ad1cb30d9f7ccda4c7854507e70429>. doi:10.1109/ICPR48806.2021.9413311.
- [3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).
- [4] A. F. Adoma, N.-M. Henry, W. Chen, Comparative analyses of bert, roberta, distilbert, and xlnet for text-based emotion recognition, in: 2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), IEEE, 2020, pp. 117–121.
- [5] T. Mikolov, M. Karafiát, L. Burget, J. Cernocký, S. Khudanpur, Recurrent neural network based language model., in: Interspeech, volume 2, Makuhari, 2010, pp. 1045–1048.
- [6] R. C. Staudemeyer, E. R. Morris, Understanding lstm—a tutorial into long short-term memory recurrent neural networks, arXiv preprint arXiv:1909.09586 (2019).
- [7] X. Qin, Z. Wu, J. Cui, T. Zhang, Y. Li, J. Luan, B. Wang, L. Wang, Bert-erc: Fine-tuning bert is enough for emotion recognition in conversation, arXiv preprint arXiv:2301.06745 (2023).
- [8] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan,

- P. Shyam, G. Sastry, A. Askill, et al., Language models are few-shot learners, *Advances in neural information processing systems* 33 (2020) 1877–1901.
- [9] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, *arXiv preprint arXiv:1810.04805* (2018).
- [10] L. Schoneveld, A. Othmani, H. Abdelkawy, Leveraging recent advances in deep learning for audio-visual emotion recognition, *Pattern Recognition Letters* 146 (2021) 1–7.
- [11] W. Zheng, H. Tang, T. S. Huang, Emotion recognition from non-frontal facial images, *Emotion Recognition: A Pattern Analysis Approach* (2015) 183–213.
- [12] A. Mircoli, G. Cimini, Automatic extraction of affective metadata from videos through emotion recognition algorithms, *Communications in Computer and Information Science* 909 (2018) 191–202. doi:10.1007/978-3-030-00063-9_19.
- [13] E. Batbaatar, M. Li, K. H. Ryu, Semantic-emotion neural network for emotion recognition from text, *IEEE access* 7 (2019) 111866–111878.
- [14] N. Alswaidan, M. E. B. Menai, A survey of state-of-the-art approaches for emotion recognition in text, *Knowledge and Information Systems* 62 (2020) 2937–2987.
- [15] I. Shahin, A. B. Nassif, S. Hamsa, Emotion recognition using hybrid gaussian mixture model and deep neural network, *IEEE access* 7 (2019) 26777–26787.
- [16] P. Ekman, An argument for basic emotions, *Cognition & emotion* 6 (1992) 169–200.
- [17] R. Plutchik, Emotions: A general psychoevolutionary theory, *Approaches to emotion* 1984 (1984) 197–219.
- [18] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, *arXiv preprint arXiv:1301.3781* (2013).
- [19] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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)*, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: <https://www.aclweb.org/anthology/N19-1423>. doi:10.18653/v1/N19-1423.
- [20] A. Chiorrini, C. Diamantini, A. Mircoli, D. Potena, Emotion and sentiment analysis of tweets using bert, in: *EDBT/ICDT Workshops, 2021*.
- [21] M. Polignano, P. Basile, M. De Gemmis, G. Semeraro, V. Basile, Alberto: Italian bert language understanding model for nlp challenging tasks based on tweets, in: *6th Italian Conference on Computational Linguistics, CLiC-it 2019, volume 2481, CEUR, 2019*, pp. 1–6.
- [22] A. Chiorrini, C. Diamantini, A. Mircoli, D. Potena, E. Storti, Emotionalbert: Emotion recognition of italian social media texts through bert, in: *2022 26th International Conference on Pattern Recognition (ICPR), 2022*, pp. 1706–1711. doi:10.1109/ICPR56361.2022.9956403.
- [23] V. Basile, M. Nissim, Sentiment analysis on italian tweets, in: *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, 2013*, pp. 100–107.
- [24] R. Sprugnoli, Multiemotions-it: A new dataset for opinion polarity and emotion analysis for italian, in: *7th Italian Conference on Computational Linguistics, CLiC-it 2020, Accademia University Press, 2020*, pp. 402–408.