

EmotivITA at EVALITA2023: Overview of the Dimensional and Multidimensional Emotion Analysis Task

Giovanni Gafà^{1,*}, Francesco Cutugno² and Marco Venuti¹

¹University of Catania, Italy

²University of Naples Federico II, Italy

Abstract

EmotivITA is the first shared task for Italian Dimensional and Multidimensional Emotion Analysis, aiming to promote research in the field of emotion detection within the Italian language. We developed an Italian dataset annotated following the dimensional model of emotions and invited participants to submit systems to predict Valence, Arousal, and Dominance associated to sentences in the *corpus*. Five runs were submitted by two teams. We present the dataset, the evaluation methodology, and the approaches of the participating systems.

Keywords

emotion analysis, emotion detection, VAD model, dataset, EmoITA, EmotivITA, Evalita 2023,

1. Introduction and Motivation

In the last two decades, the analysis of emotions that people express in texts has become an essential area in Natural Language Processing (NLP). Such an interest springs from the awareness of the crucial role feelings have in our cognition: being able to detect and eventually simulate them could be a fundamental step to produce human-like forms of artificial intelligence [1]. For a review on possible applications of Emotion Analysis (EA), ranging from stock market predictions to the management of catastrophic events, see for example [2].

Taking into account the somewhat uncertain terminology about human feelings occasionally found in the literature (see below), we start by defining some terms. Adopting a well known typology of affective states by Scherer [3, pp. 140–141], we use the word ‘emotion’ to refer to a “relatively brief episode of synchronized responses by all or most organismic subsystems to the evaluation of an external or internal event as being of major significance”, whereas ‘sentiments’, like Scherer’s ‘attitudes’, are “relatively enduring, affectively colored beliefs, preferences, and predispositions toward objects or persons”.

Sentiment analysis has been a major interest for computational linguistics for a long time, and, over the years, it moved from the prediction of the semantic polarity towards more fine-grained modeling, as is the case in *Aspect-based Sentiment Analysis* [4] and *Stance Detection* [5]; similar studies have been conducted on Italian

texts as well [6, 7].

Recently, EA started receiving more and more attention as well. Several models of emotion proposed in psychology have been used in NLP, either categorical or dimensional. The former consider feelings as discrete, and usually identify a small set of basic emotions upon which other, more subtle and complex affective states are built; the widely adopted model conceived by Ekman [8], for instance, proposes six fundamental emotions. The latter, on the contrary, describes emotions by combining a limited number of independent dimensions in a real-valued vector space. The model proposed by Russell and Mehrabian [9], probably the best-known, recognizes three dimensions: *Valence* (measuring pleasure or displeasure), *Arousal* (degree of excitement or calm), and *Dominance* (level of control over the situation) – the VAD model.

Categorical models have some advantages over dimensional ones, as they allow the identification of several emotions in the same input and usually have simpler interpretations. Nevertheless, they have been criticized for their use of culture and language specific labels [10]; besides, different categorical models adopt different sets of emotions, making it difficult to compare studies. Concerning dimensional models, the independence of the three dimensions is yet to be ascertained [11, 12]; however, dimensional models allow easier comparisons between emotions and can describe feelings that are difficult to label.


At SemEval, the most renowned evaluation campaign of NLP, the first shared task concerning emotion detection (for three languages: English, Arabic and Spanish) was proposed in 2018 [13]. Building on earlier works, a 22,000 tweet dataset was annotated for many different affect states, following both the categorical and dimensional models of emotions (limited to the Valence dimen-

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

*Corresponding author.

✉ giovanni.gafa@phd.unict.it (G. Gafà); cutugno@unina.it (F. Cutugno); marco.venuti@unict.it (M. Venuti)

© 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)

sion¹); the sub-tasks involved emotion classification and emotion regression.

Another task of emotion classification was proposed at SemEval 2019 [14], this time leveraging a dataset containing roughly 3,000 short conversations annotated for the presence of four emotions; the purpose was to study and exploit the role of context in facilitating emotion detection.

Anyway, EA has not yet received in Italy the same amount of interest it gained at the international level. This is probably due to the lack of resources annotated for emotions. After some investigations, we could find just a few *lexica* [15, 16, 17]; some are not open to the public [18] or are quite specific in scope [19]; others are the result of automatic translations from English of existing vocabularies, and have not been re-annotated by Italian speakers [20, 21]. This situation worsens when it comes to datasets, where to the best of our knowledge only domain-specific resources are available [22, 23, 24]. Another dataset [25] has been proposed at Evalita 2023 [26], containing social media messages about TV shows, TV series, music videos, and advertisements, which had been labeled following the Plutchik model of emotions [27].

As we tried to outline, existing datasets for EA in Italian are scarce and quite specialized. Moreover, the emotion formats used for annotating the *corpora* are uniquely categorical. Nevertheless, dimensional models are receiving increasing attention in tasks of emotion detection [28, 29]. By proposing the EmotivITA shared task at the Evalita 2023 evaluation campaign, we aim at providing a new, general-purpose resource for EA in Italian, with labeling provided by Italian speakers, EmoITA: a dataset composed with a genre and domain-balanced selection of more than 10,000 written sentences, annotated following the dimensional model of emotions; on the other hand, we intend to promote dimensional and multidimensional EA in Italian.

The rest of the paper is organized as follows: Section 2 provides a definition of the task; Section 3 describes the dataset made available to participants, and the process of its creation; Section 4 details the official evaluation measures; Section 5 reports the results obtained by participating teams; Section 6 discusses the results; in Section 7 we draw some conclusions on the outcomes of the task.

2. Definition of the task

The EmotivITA shared task consists of automatically annotating for emotions in the VAD model a collection of written sentences from a genre-balanced dataset translated into Italian. More specifically, the task has been

¹As a case in point of inaccuracy when dealing with emotion-related terms, Valence was regarded as an equivalent of ‘sentiment’ throughout the study.

organized into two sub-tasks whose results will be evaluated separately:

- **Sub-task A: Dimensional emotion regression.** Prediction of Valence, Arousal, and Dominance values based on a set of Italian sentences and annotations, using only the target annotated dimension for training – so, for instance, when predicting Valence participant systems may only use Valence values annotated in the dataset for training; the same holds for Arousal and Dominance.
- **Sub-task B: Multidimensional emotion regression.** Prediction of Valence, Arousal, and Dominance values based on a set of Italian sentences and annotations, using all mentioned dimensions for training – so participant systems should determine Valence, Arousal, and Dominance simultaneously, using values from the three dimensions for training.

Both sub-tasks are regression problems, so participating teams were asked to provide in the output the sentence id and three real numbers between 1 and 5, relative to the three predicted dimensions. Sub-task B intends to study and exploit potential correlations between Valence, Arousal, and Dominance, which have been discussed in the literature (see § 1).

Participants could carry out either both sub-tasks or only one of them, even if participation in sub-task A was strongly recommended, in order to have a common basis for comparison. Each participating team was allowed to submit a maximum of 2 runs for each sub-task. All runs could be produced according to the ‘constrained’ or ‘unconstrained’ modality (or both); however, we asked to specify the type of run. In constrained modality, only annotated data distributed by the organizers could be used for training and tuning the systems. Other linguistic resources (e.g., word embeddings and lexicons) were instead allowed. In unconstrained modality, annotated external data could also be employed and had to be described in the system reports.

3. Dataset

As mentioned above, the data released for the shared task derive from the Italian translation of an existing dataset, EmoBank [30]. EmoBank is the largest genre-balanced English dataset annotated employing the VAD model of emotions. As shown in Table 1, it mainly consists of the *MASC: Manually Annotated Sub-Corpus of the American National Corpus* [31], with roughly 10% of the sentences coming from the dataset of SemEval-2007 Task 14 [32].

The 10,062 sentences were originally annotated by English native speakers according to two different perspec-

Table 1
Genre distribution of the EmoBank corpus.

Corpus	Domain	# Sentences
SemEval-2007	news headlines	1,192
MASC	blogs	1,336
	essays	1,135
	fiction	2,753
	letters	1,413
	newspapers	1,314
	travel guides	919
Total		10,062

tives: the emotion they felt the writer meant to express, and the emotion evoked in an average reader.

At first, the Italian version of the dataset was studied as part of a Master’s degree thesis discussed in 2022 at the Department of Humanities of the University of Catania. In this context, the sentences were initially translated automatically to Italian using the neural machine translation service offered by Microsoft Azure. As we were not satisfied with the results, a manual revision was performed splitting the *corpus* evenly between nine Italian native speakers, researchers in linguistics affiliated with Interdepartmental Research Center Urban/Eco at the University of Naples Federico II.

We also conducted a pilot study asking two of the participants to independently annotate VAD values from the reader’s perspective for a small sample of sentences (150). We chose the reader’s perspective because, according to Buechel and Hahn, it yields better inter-annotator agreement (IAA). For annotation, we used the Self-Assessment Manikin (SAM), a pictographic scale to assess emotional response [33, 34] already adopted for EmoBank. SAM consists of three sets of anthropomorphic cartoons displaying differences in Valence, Arousal, and Dominance values, respectively as shown in Figure 1.

We asked participants to attribute a value between 1 (minimum Valence, Arousal, and Dominance) and 5 (maximum Valence, Arousal, and Dominance), with 4 intermediate steps of 0.5. This results in a 9-point scale like the one originally proposed by Bradley and Lang (Buechel and Hahn preferred a 5-point scale). Instructions were adapted from those used for EmoBank and are available for further analysis upon request.

To measure IAA we used Pearson’s correlation coefficient (r) and Mean Absolute Error (MAE), as other metrics like Cohen’s k are not designed for scale variables (see § 4). We obtained encouraging scores in both measures for all three dimensions, with an average of 0.593 for r , indicating a large effect (see Table 2). Therefore, we decided to ask all participants to annotate the remaining sentences individually (one annotator per sentence). We then used the new labeling to fine-tune several models of transformers for dimensional EA, but the scores were

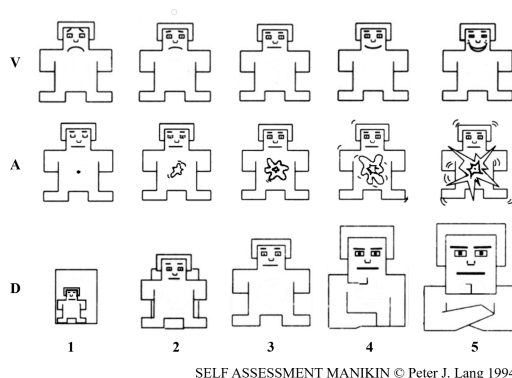


Figure 1: The SAM scales for VAD values. Dimensions (Valence, Arousal and Dominance) are reported in rows, values (from 1 to 5) in columns. Copyright of SAM by Peter J. Lang 1994.

Table 2
IAA for the three dimensions in the pilot study.

	V	A	D	Average
r	0.794	0.552	0.676	0.593
MAE	0.357	0.900	0.583	0.613

significantly worse than those obtained with the original values from the EmoBank dataset (MAE was between 2 and 3 times higher, r for Valence and Arousal was respectively 33% and 13% lower). This was probably due to the lack of consistency from having a single annotation for a sentence.

Moreover, we reviewed the manual revisions of the translations and found that, in at least half of the cases, the quality was still poor, either because the translated sentence did not feel natural in Italian or because it contained some kind of error.

To produce EmoITA, we resolved to start over the entire process, only keeping the approximately 5,000 translations we considered good enough. This time, we chose 16 students from the Master’s Degree in Foreign Languages at the University of Catania. All of them are Italian native speakers and are specializing in English. The sentences were split among the participants: we asked to revise the 5,000 translations we kept from previous work and to propose new translations for the rest of the *corpus*. The same group of subjects also labeled each Italian sentence, and we took care never to ask a participant to annotate a sentence he had translated. Overall, we obtained 7 different annotations for each sentence, and we judge the quality of translation is now satisfactory if not perfect.

To evaluate the annotations, we proceeded similarly to the original EmoBank study: we calculated r and MAE between each individual series of annotations and the

Table 3
IAA for the three dimensions in EmoITA.

	V	A	D	Average
r	0.702	0.507	0.535	0.581
MAE	0.496	0.536	0.489	0.507

Table 4
Annotation examples from the development dataset.

id	text	V	A	D
260	Jet si capovolge durante una tormenta, nessun morto.	3.33	4.17	3
261	Certo, risposi.	3.83	3.17	3.83

aggregated values in EmoITA, and then averaged those values for each dimension (see Table 3).

The values of r indicate a large effect in every dimension, particularly for Valence. Correlation is a little higher in Dominance than in Arousal, as per our pilot study: this is somewhat unusual, as in most research we analyzed regarding the English language the opposite is true. MAE is not as good, but still acceptable (10% of the 1-5 scale). Overall, scores are in line with those of EmoBank ($r=0.634$ and $MAE=0.386$, on average). They could probably get better analyzing outliers and excluding some of the annotations whose disagreement is particularly strong, a process we have not yet started at this time.

For the shared task the dataset was randomly split into a development and a test set of 8,000 and 2,062 sentences respectively (79.5% and 20.5%), taking care to preserve the genre distribution in the *corpus* (with a 1% tolerance). The development set was provided as a UTF-8, CSV comma-separated file, reporting the following fields:

id, text, V, A, D

where:

1. ‘id’ denotes the unique identifier of the sentence
2. ‘text’ denotes the text of the sentence
3. ‘V’ denotes the average Valence value annotated for the sentence.
4. ‘A’ denotes the average Arousal value annotated for the sentence.
5. ‘D’ denotes the average Dominance value annotated for the sentence.

See Table 4 for a couple of examples.

The test set followed the same format, but labels for Valence, Arousal and Domination were not provided.

4. Evaluation Measures

The two sub-tasks are evaluated separately comparing results obtained by participant systems with the gold

standard annotations of the test set. Both constrained and unconstrained runs for a sub-task are reported in the same ranking, but we specify the type of the run.

Evaluation metrics for both sub-tasks are the standard metrics known in the literature for emotion regression that we already mentioned throughout this paper: we measure IAA based on r and MAE . The first metric estimates linear dependence between two series of data points: $\mathbf{x} = x_1, \dots, x_n$ and $\mathbf{y} = y_1, \dots, y_n$. In our case, \mathbf{x} corresponds to the values annotated in our dataset for each dimension and \mathbf{y} to those predicted by participant systems. The formula for r is as follows:

$$r(\mathbf{x}, \mathbf{y}) := \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (1)$$

where \bar{x} and \bar{y} are respectively the mean value of \mathbf{x} and \mathbf{y} .

MAE is a measure of errors between a couple of observations describing the same phenomenon (in this case the annotated values of a certain emotional dimension in the dataset, and those predicted). The formula for MAE is as follows:

$$MAE(\mathbf{x}, \mathbf{y}) := \frac{1}{n} \sum_{i=1}^n |x_i - y_i| \quad (2)$$

The baselines for both sub-tasks have been built fine-tuning to a regression a BERT model available on HuggingFace², with a learning rate of 1e-05.

5. Results

We received submissions from two teams. Both of them participated to sub-task B, and only one to sub-task A. In total, 5 runs were submitted, constrained and unconstrained. In Table 5 we report the results for r and MAE in sub-task A, in Table 6 those relative to sub-task B, along with our baselines. We appended a suffix to distinguish the ID of the submitted run and another one to identify constrained (‘_C_’) and unconstrained (‘_U_’) runs.

Regarding sub-task A, the ISTC-CNR team obtained the best r score in the Valence dimension with his second run. Anyway, our baseline had better results in every other dimension and metric.

Concerning sub-task B, the team extremITA achieved the best results in all metrics and dimensions with their second run, with the exception of Arousal and Dominance’s r , where our baseline performed slightly better.

6. Discussion

The teams of the EmotivITA challenge were invited to describe their solution in a technical report; in this section

²<https://huggingface.co/dbmdz/bert-base-italian-xxl-uncased>, last access 06-20-2023.

Table 5

Results of the submissions for sub-task A.

System	V r	A r	D r	V MAE	A MAE	D MAE
ISTC-CNR_1_C	0.800	0.593	0.609	0.393	0.432	0.399
ISTC-CNR_2_C	0.809	0.587	0.624	0.382	0.433	0.387
baseline	0.807	0.643	0.643	0.313	0.321	0.285

Table 6

Results of the submissions for sub-task B.

System	V r	A r	D r	V MAE	A MAE	D MAE
ISTC-CNR_1_C	0.800	0.594	0.623	0.400	0.421	0.367
extremITA_1_U	0.708	0.430	0.548	0.327	0.395	0.297
extremITA_2_U	0.811	0.633	0.630	0.272	0.296	0.266
baseline	0.811	0.652	0.654	0.859	0.859	0.859

we compare participant systems based on their architectures.

The ISTC-CNR team proposed a method based on Natural Language Inference (NLI). More specifically, they used a multilingual MNLI-XML-RoBERTa model grounded on XML-RoBERTa [35], which was fine-tuned on a version of the MNLI dataset [36] automatically translated to Italian. The model was adapted for the regression task replacing its last linear layer. During training, sentences from the EmoITA dataset were used as premises. Then, for sub-task A, three different models were conceived, with three different prompts acting as hypotheses for the NLI process and targeting the VAD dimensions. The prompt for Valence was “quanta positività esprime la frase?” (how much positivity does the sentence convey?), the one for Arousal “quanto è eccitante la frase?” (how exciting is the sentence?) and the one for Domination “quanto è controllata l’emozione” (how controlled is the emotion?). For sub-task B, a single model was used adopting the prompt “valence, arousal, dominance dell’emozione?” (valence, arousal, and dominance of the emotion?). The two runs submitted for sub-task A differ in that the first one only utilized 99% of the training set made available, while the second one utilized it entirely. As we can see in Table 5, the results were better with this last configuration. The only run submitted by the team for sub-task B exploited the entire training set. All runs were produced according to the constrained modality.

The extremITA team only participated to sub-task B, with two unconstrained runs. Both their models were trained on the union of all the datasets in the shared tasks at EVALITA 2023. The first one adopts an Encoder-Decoder architecture based on IT5 [37], a T5 model [38] pre-trained on Italian texts. The model was fine-tuned concatenating the name of the shared task as a prefix, followed by an input sentence from the EmoITA dataset. The output, in the case of the EmotivITA task, was constituted by the predicted VAD values. A similar approach

was used for every task in EVALITA 2023. The second architecture is a Decoder that adopts instruction-tuning, based on a large language model, the LLaMA [39]. The model was trained using Low-Rank Adaptation [40] on Italian translations of the instructions originally developed for Alpaca [41], which also builds on LLaMA. It was then fine-tuned using instructions specific to the addressed EVALITA task. In the case of EmotivITA sub-task B, the sentence from the EmoITA dataset was paired with a prompt in the form of the instruction: “Scrivi quanta valenza è espressa in questo testo su una scala da 1 a 5, seguito da quanto stimolo è espresso in questo testo su una scala da 1 a 5, seguito da quanto controllo è espresso in questo testo in una scala da 1 a 5” (Rate how much valence is expressed in this text on a scale from 1 to 5, followed by how much arousal is expressed in this text on a scale from 1 to 5, followed by how much dominance is expressed in this text on a scale from 1 to 5). This second model obtained generally better performance than the first one as showcased in Table 6, but it also demanded 144 hours of training (on the entire EVALITA dataset), whereas the one based on IT5 only required 12 hours.

Quite interestingly the model proposed by the ISTC-CNR team and the second one proposed by the extremITA team both leverage prompting in natural language and no task-specific architectural designs (with the exception of the replacement of the last layer in the MNLI-XML-RoBERTa model), proving the efficacy of this approach. On the other hand, one could argue that the main limitations of the ISTC-CNR method was precisely the chosen prompts, as concepts like Valence, Arousal and Dominance are not easy to describe. When evaluating the extremITA proposal, instead, one could wonder about the sustainability of a 144 hours training process.

Anyway, we observe that the baselines obtained fine-tuning the BERT model were not outperformed by the proposed systems: maybe the upper limit for the regression problem with such a large dataset as EmoITA has

been reached, at least for the moment. It is also worth mentioning that the scores are in line with those of the study representing the state-of-the-art [42] for the original English dataset, EmoBank, that obtained values of 0.838, 0.573 and 0.536 for r in the three dimensions.

One last remark is due; neither team explored the possible relations between the three emotion dimensions, which was actually one of the purposes of sub-task B, and remains as a subject for future studies.

7. Conclusion

We presented the first shared task on Dimensional and Multidimensional Emotion Analysis for Italian and discussed the development of the first dedicated Italian dataset, EmoITA, based on the VAD model. EmoITA was obtained by manual translation and annotation of the EmoBank dataset, performed by Italian native speakers. The participating systems leveraged NLI, the Encoder-Decoder architecture and Large Language Models to address the regression problems, obtaining results that are similar to those of the state-of-the-art for the English counterpart of the dataset.

We hope that the proposal of our task and the availability of a new Italian dataset for EA will foster studies in this relevant field of NLP. In this spirit, the development and test set, as well as the complete dataset (licensed under CC-BY-SA 4.0), the script used for the baselines and for evaluation will be made available to the public soon; more details on EmotivITA can be found on the task website³.

References

- [1] R. W. Picard, *Affective computing*, MIT Press, Cambridge, Mass, 1997.
- [2] M. Akhtar, D. Ghosal, A. Ekbal, P. Bhattacharyya, S. Kurohashi, All-in-one: Emotion, sentiment and intensity prediction using a multi-task ensemble framework, *IEEE Transactions on Affective Computing* 13 (2022) 285–297. doi:10.1109/TAFFC.2019.2926724.
- [3] K. R. Scherer, *Psychological Models of Emotion*, in: J. C. Borod (Ed.), *The Neuropsychology of Emotion*, Oxford University Press, New York, 2000, pp. 137–162.
- [4] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, *SemEval-2014 task 4: Aspect based sentiment analysis*, in: *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, Association for Computational Linguistics, Dublin, Ireland, 2014, pp. 27–35. URL: <https://aclanthology.org/S14-2004>. doi:10.3115/v1/S14-2004.
- [5] S. Mohammad, S. Kiritchenko, P. Sobhani, X. Zhu, C. Cherry, *SemEval-2016 task 6: Detecting stance in tweets*, in: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, Association for Computational Linguistics, San Diego, California, 2016, pp. 31–41. URL: <https://aclanthology.org/S16-1003>. doi:10.18653/v1/S16-1003.
- [6] A. Cignarella, M. Lai, C. Bosco, V. Patti, P. Rosso, *SardiStance @ EVALITA2020: Overview of the Task on Stance Detection in Italian Tweets*, in: V. Basile, D. Croce, M. Maro, L. C. Passaro (Eds.), *EVALITA Evaluation of NLP and Speech Tools for Italian - December 17th, 2020*, Accademia University Press, 2020, pp. 177–186. URL: <http://books.openedition.org/aaccademia/7084>. doi:10.4000/books.aaccademia.7084.
- [7] P. Basile, D. Croce, V. Basile, M. Polignano, *Overview of the EVALITA 2018 Aspect-based Sentiment Analysis task (ABSITA)*, in: T. Caselli, N. Novielli, V. Patti, P. Rosso (Eds.), *EVALITA Evaluation of NLP and Speech Tools for Italian*, Accademia University Press, 2018, pp. 10–16. URL: <http://books.openedition.org/aaccademia/4451>. doi:10.4000/books.aaccademia.4451.
- [8] P. Ekman, *Basic Emotions*, in: T. Dalgleish, M. J. Power (Eds.), *Handbook of Cognition and Emotion*, John Wiley & Sons, Ltd, Chichester, UK, 2005, pp. 45–60. URL: <https://onlinelibrary.wiley.com/doi/10.1002/0470013494.ch3>. doi:10.1002/0470013494.ch3.
- [9] J. A. Russell, A. Mehrabian, *Evidence for a three-factor theory of emotions*, *Journal of Research in Personality* 11 (1977) 273–294. URL: <https://linkinghub.elsevier.com/retrieve/pii/009265667790037X>. doi:10.1016/0092-6566(77)90037-X.
- [10] E. Cambria, A. Livingstone, A. Hussain, *The hourglass of emotions*, in: A. Esposito, A. M. Esposito, A. Vinciarelli, R. Hoffmann, V. C. Müller (Eds.), *Cognitive Behavioural Systems*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, pp. 144–157.
- [11] P. Kuppens, F. Tuerlinckx, J. A. Russell, L. F. Barrett, *The relation between valence and arousal in subjective experience*, *Psychological Bulletin* 139 (2013) 917–940. URL: <http://doi.org/getdoi.cfm?doi=10.1037/a0030811>. doi:10.1037/a0030811.
- [12] A. B. Warriner, V. Kuperman, M. Brysbaert, *Norms of valence, arousal, and dominance for 13,915 English lemmas*, *Behavior Research Methods* 45 (2013) 1191–1207. URL: <http://link.springer.com/10.3758/s13428-012-0314-x>. doi:10.

³Repository: <https://github.com/GiovanniGafa/EmoITA>. Website: <https://sites.google.com/view/emotivita>

- 3758/s13428-012-0314-x.
- [13] S. Mohammad, F. Bravo-Marquez, M. Salameh, S. Kiritchenko, SemEval-2018 task 1: Affect in tweets, in: Proceedings of the 12th International Workshop on Semantic Evaluation, Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 1–17. URL: <https://aclanthology.org/S18-1001>. doi:10.18653/v1/S18-1001.
- [14] A. Chatterjee, K. N. Narahari, M. Joshi, P. Agrawal, SemEval-2019 task 3: EmoContext contextual emotion detection in text, in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 39–48. URL: <https://aclanthology.org/S19-2005>. doi:10.18653/v1/S19-2005.
- [15] O. Araque, L. Gatti, J. Staiano, M. Guerini, DepecheMood++: A Bilingual Emotion Lexicon Built Through Simple Yet Powerful Techniques, IEEE Transactions on Affective Computing 13 (2022) 496–507. URL: <https://ieeexplore.ieee.org/document/8798675/>. doi:10.1109/TAFFC.2019.2934444.
- [16] M. Montefinese, E. Ambrosini, B. Fairfield, N. Mammarella, The adaptation of the Affective Norms for English Words (ANEW) for Italian, Behavior Research Methods 46 (2014) 887–903. URL: <https://link.springer.com/10.3758/s13428-013-0405-3>. doi:10.3758/s13428-013-0405-3.
- [17] L. Passaro, L. Pollacci, A. Lenci, ItEM: A Vector Space Model to Bootstrap an Italian Emotive Lexicon, Second Italian Conference on Computational Linguistics CLiC-it 2015 II (2015).
- [18] A. Bolioli, F. Salamino, V. Porzionato, Social Media Monitoring in Real Life with Blogmeter Platform, in: C. Battaglini, C. Bosco, E. Cambria, R. Damiano, V. Patti, P. Rosso (Eds.), Proceedings of the First International Workshop on Emotion and Sentiment in Social and Expressive Media: approaches and perspectives from AI (ESSEM 2013) A workshop of the XIII International Conference of the Italian Association for Artificial Intelligence (AI*IA 2013), Turin, Italy, December 3, 2013, volume 1096 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2013, pp. 156–163. URL: <http://ceur-ws.org/Vol-1096/paper12.pdf>.
- [19] E. Borelli, D. Crepaldi, C. A. Porro, C. Cacciari, The psycholinguistic and affective structure of words conveying pain, PLOS ONE 13 (2018) e0199658. URL: <https://dx.plos.org/10.1371/journal.pone.0199658>. doi:10.1371/journal.pone.0199658.
- [20] S. M. Mohammad, P. D. Turney, Crowdsourcing a word-emotion association lexicon, Computational Intelligence 29 (2013) 436–465.
- [21] S. M. Mohammad, Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words, in: Proceedings of The Annual Conference of the Association for Computational Linguistics (ACL), Melbourne, Australia, 2018.
- [22] Celli, Fabio, Riccardi, Giuseppe, Ghosh, Aridam, CorEA: Italian news corpus with emotions and agreement, in: Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 and of the Fourth International Workshop EVALITA 2014 9-11 December 2014, Pisa, PISA UNIVERSITY PRESS, 2014. URL: <http://clit2014.fileli.unipi.it/proceedings/Proceedings-CLiC-it-2014.pdf>. doi:10.12871/CLICIT2014120.
- [23] Z. Shibingfeng, F. Francesco, G. Federico, B.-C. Alberto, B. Paolo, P. Angelo, AriEmozione2.0, 2022. URL: <https://zenodo.org/record/7097913>. doi:10.5281/ZENODO.7097913.
- [24] R. Sprugnoli, MultiEmotions-It: a New Dataset for Opinion Polarity and Emotion Analysis for Italian, in: J. Monti, F. dell’Orletta, F. Tamburini (Eds.), Proceedings of the Seventh Italian Conference on Computational Linguistics, CLiC-it 2020, Bologna, Italy, March 1-3, 2021, volume 2769 of *CEUR Workshop Proceedings*, CEUR-WS.org, Torino, 2020. URL: http://ceur-ws.org/Vol-2769/paper_08.pdf. doi:10.4000/books.aaccademia.8910.
- [25] O. Araque, S. Frenda, D. Nozza, V. Patti, R. Sprugnoli, Emit at evalita2023: Overview of the categorical emotion detection in italian social media task, in: M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi (Eds.), Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.
- [26] 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.
- [27] R. Plutchik, A General Psychoevolutionary Theory of Emotion, in: Theories of Emotion, Elsevier, 1980, pp. 3–33. URL: <https://linkinghub.elsevier.com/retrieve/pii/B9780125587013500077>. doi:10.1016/B978-0-12-558701-3.50007-7.
- [28] R. Mukherjee, A. Naik, S. Poddar, S. Dasgupta, N. Ganguly, Understanding the role of affect dimensions in detecting emotions from tweets: A multi-task approach, CoRR abs/2105.03983 (2021). URL: <https://arxiv.org/abs/2105.03983>. arXiv:2105.03983.
- [29] J. Wang, L.-C. Yu, K. R. Lai, X. Zhang, Dimensional sentiment analysis using a regional CNN-LSTM model, in: Proceedings of the 54th Annual Meet-

- ing of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Berlin, Germany, 2016, pp. 225–230. URL: <https://aclanthology.org/P16-2037>. doi:10.18653/v1/P16-2037.
- [30] S. Buechel, U. Hahn, EmoBank: Studying the Impact of Annotation Perspective and Representation Format on Dimensional Emotion Analysis, in: M. Lapata, P. Blunsom, A. Koller (Eds.), Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers, Association for Computational Linguistics, 2017, pp. 578–585. URL: <http://aclweb.org/anthology/E17-2092>. doi:10.18653/v1/E17-2092.
- [31] N. Ide, C. Baker, C. Fellbaum, C. Fillmore, R. Passonneau, MASC: the manually annotated sub-corpus of American English, in: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), European Language Resources Association (ELRA), Marrakech, Morocco, 2008. URL: http://www.lrec-conf.org/proceedings/lrec2008/pdf/617_paper.pdf.
- [32] C. Strapparava, R. Mihalcea, SemEval-2007 task 14: Affective text, in: Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), Association for Computational Linguistics, Prague, Czech Republic, 2007, pp. 70–74. URL: <https://aclanthology.org/S07-1013>.
- [33] M. M. Bradley, P. J. Lang, Measuring emotion: The self-assessment manikin and the semantic differential, *Journal of Behavior Therapy and Experimental Psychiatry* 25 (1994) 49–59. URL: <https://linkinghub.elsevier.com/retrieve/pii/0005791694900639>. doi:10.1016/0005-7916(94)90063-9.
- [34] P. J. Lang, Behavioral treatment and bio-behavioral assessment: Computer applications, in: J. B. Sidowski, J. H. Johnson, T. A. Williams (Eds.), *Technology in mental health care delivery systems*, Norwood, NJ: Ablex Publishing, 1980, pp. 119–137.
- [35] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, *CoRR* abs/1911.02116 (2019). URL: <http://arxiv.org/abs/1911.02116>. arXiv:1911.02116.
- [36] A. Williams, N. Nangia, S. Bowman, A broad-coverage challenge corpus for sentence understanding through inference, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 1112–1122. URL: <https://aclanthology.org/N18-1101>. doi:10.18653/v1/N18-1101.
- [37] G. Sarti, M. Nissim, IT5: Large-scale text-to-text pretraining for italian language understanding and generation, *ArXiv preprint 2203.03759* (2022). URL: <https://arxiv.org/abs/2203.03759>.
- [38] 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.
- [39] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, Llama: Open and efficient foundation language models, 2023. arXiv:2302.13971.
- [40] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, 2021. arXiv:2106.09685.
- [41] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, T. B. Hashimoto, Stanford alpaca: An instruction-following llama model, https://github.com/tatsu-lab/stanford_alpaca, 2023.
- [42] S. Park, J. Kim, S. Ye, J. Jeon, H. Y. Park, A. Oh, Dimensional emotion detection from categorical emotion, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 4367–4380. URL: <https://aclanthology.org/2021.emnlp-main.358>. doi:10.18653/v1/2021.emnlp-main.358.