EMit at EVALITA 2023: Overview of the Categorical Emotion Detection in Italian Social Media Task*

Oscar Araque[1,*,†], Simona Frenda[2,3,†], Rachele Sprugnoli[4,†], Debora Nozza[5,†] and Viviana Patti[2,†]

[1]Intelligent Systems Group, Universidad Politécnica de Madrid, Spain
[2]Università degli Studi di Torino, Italy
[3]aequa-tech srl, Turin, Italy
[4]Università di Parma, Italy
[5]Università Bocconi, Italy

Abstract

The Emotions in Italian (EMit) task is the first edition of a shared task on emotion analysis and opinion mining in Italian messages at EVALITA 2023. EMit presents two subtasks: (i) Subtask A, that consists in an emotion detection challenge, and (ii) Subtask B, that introduces a novel problem of target detection of the expressed emotion. Additionally, EMit challenges systems with a thorough in-domain and out-of-domain evaluation, probing the generalization capabilities of the submitted solutions. In general, 4 teams have participated in Subtask A, achieving a macro-averaged f-score of 0.6028 and 0.4977 in the in-domain and out-of-domain sets, respectively. In Subtask B a team has participated, obtaining 0.6459 in the in-domain set and 0.3223 in the out-of-domain set as macro-averaged f-scores. The obtained results indicate that further work needs to be done to solve the task, opening new avenues of research.

Keywords

Emotion detection, Emotion target detection, User-generated contents, Sentiment Analysis

1. Introduction and Motivations

The detection of emotions in texts has a long history in international evaluation campaigns but has never been addressed in EVALITA where the only shared task to deal with emotions was about emotional speech recognition systems [1]. The *Affective Text* shared task at SemEval 2007 was the first one to propose the classification of newspaper headlines according to 6 emotions: anger, disgust, fear, joy, sadness, surprise [2]. Then, starting from 2017, this type of evaluation has become very frequent with a particular attention to the processing of tweets and dialogues. For example, *Affect in Tweets* at SemEval 2018 [3] included a subtask about the multilabel detection of 11 emotions in tweets written in English, Arabic, and Spanish whereas the *Emotion Detection* task at TASS 2020 [4] and EmoEvalEs at IberLEF 2021 [5] were only on Spanish tweets [1]. Instead, *EmoContext* at SemEval 2019 [6] and *EmotionX* at the SocialNLP workshop in 2018 [7] and 2019 focused on the emotion classification of dialogues in English. Last year, the *Emotion Classification* shared task at WASSA 2022 dealt with a different genre of text proposing the classification of emotions in essays written in reaction to news articles [8].

In this context, the **EMit (Emotions in Italian)** task[2] aims at providing the first evaluation framework for emotion detection in Italian texts at EVALITA [9], offering novel annotated data available to the community that will foster future research. EMit tackles a comprehensive emotion model that is complemented with additional annotations regarding the scope of opinions.

2. Task Description

EMit is organized according to two subtasks, thus offering participants different perspectives on opinion analysis:

- **Subtask A**
  Emotion Detection (Main Task) The main proposed subtask is the detection of *emotions* in social media messages about TV shows and series

[1]https://competitions.codalab.org/competitions/28682
[2]Task website: http://www.di.unito.it/~tutreeb/emit23/index.html, task repository: https://github.com/oaraque/emit

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT
*Corresponding author.
o.araque@upm.es (O. Araque); simona.frenda@unito.it (S. Frenda); rachele.sprugnoli@unipr.it (R. Sprugnoli); debora.nozza@unibocconi.it (D. Nozza); viviana.patti@unito.it (V. Patti)
https://gsi.upm.es/oaraque/ (O. Araque); http://www.di.unito.it/~frenda/ (S. Frenda); https://personale.unipr.it/it/ugovdocenti/person/236480 (R. Sprugnoli); https://deboranozza.com/ (D. Nozza); https://www.unito.it/persone/vpatti (V. Patti)
0000-0003-3224-0001 (O. Araque); 0000-0002-6215-3374 (S. Frenda); 0000-0001-6861-5595 (R. Sprugnoli); 0000-0002-7998-2267 (D. Nozza); 0000-0001-5991-370X (V. Patti)

emitted by RAI (Radiotelevisione italiana, the national public broadcasting company of Italy), music videos and advertisements.
Given a message, the system decides the emotions expressed in the message or the absence of emotions.

- **Subtask B**
  Target Detection The second subtask is about the detection of the target addressed by the author of the message: the *topic* or the *direction*. In each text, it is indicated whether this refers to what the broadcast is about (the *topic*) or whether it refers to something that is under control of the broadcast itself (*direction*). When the target of the post is the *topic*, this means that the text addresses topics such as events, issues discussed in the TV episode/music video/advertisements, or invited guests of a TV show. On the other hand, the target encoded as *direction* implies that the message describes the specific directors of the shows/series, the showman/artists, fixed guests in the TV shows, reporters, or the show/series/-music video/advertisements as such.
  Given a message, the system decides if the target of the message is related to topic, direction, both or none of the two.

Both subtasks are designed as **multilabel problems of classification**. In this way, participating systems are required to provide as output the *id* of the message and all the predicted labels contained in it. It is worth mentioning that in Subtask A, the message may be classified as neutral, or expressing one or more emotions. Thus, the provided labels are: `neutral` when the message does not express any emotion, the 8 main emotions defined by Plutchik in [10] (`anger`, `anticipation`, `disgust`, `fear`, `joy`, `sadness`, `surprise`, `trust`), and the additional label `love` that is one of the primary dyads in the Plutchik's wheel of emotions, being a combination of *joy* and *trust*. Therefore, a total of 10 labels are used for Subtask A. In Subtask B the message can be classified as addressing the topic, the direction, or both or neither, thus the provided labels are: `topic` and `direction`.

Considering the specific attention on the entertainment sector, we designed Subtask B particularly on the events and players involved in such contents and in their creation. Indeed, the combination of the two subtasks allows going beyond the simple detection of emotions, identifying also if the target of the affective comments about TV programs is related to the *topic* or to issues under control of the broadcasting company (the *direction*). Such finer grained information can be of great importance in real application domains, for artists or broadcasters in the evaluation of the contents delivered, when the analysis of emotions in social media is used as a social signal of emotional reactions of Italian television audience. In other words, this would lead to the development of an *Auditel of emotions*.

3. Datasets

In order to evaluate the robustness of the models proposed by participants, in EMit we release two different test sets: (i) the *in-domain* dataset, which including tweets of the same textual genre and subjects of the training set, and (ii) an additional *out-of-domain* set that is composed of social text of different genres and subjects[3]. In this way, we offer to participants a cross-domain evaluation setting for both subtasks A and B. Table 1 summarizes the size and distribution of the datasets used in EMit 2023.

Learning Set	Dataset	Total (approx.)
Subtask A		
Train	In-domain	5,966
Test 1	In-domain	1,000
Test 2	Out-of-domain	1,000
Subtask B		
Train	In-domain	5,966
Test 1	In-domain	1,000
Test 2	Out-of-domain	1,000

Table 1
EMit 2023 datasets and their distribution in subtasks.

Dataset for in-domain evaluation.
This dataset is obtained from Twitter and it is composed of 6,966 tweets that discuss programs by the Italian RAI TV station. Such messages have been grouped in almost 5 set, each set annotated by three different annotators (for a total of 15 annotators) with a multi-layered annotation scheme. As described, the *emotion* layer consists of 10 labels: Plutchik's emotions, `love` and `neutral`. These emotion annotations are used for running Subtask A.

The emotion labels are non-exclusive, thus a certain tweet can be annotated with one or more emotions, or even solely as neutral , as shown in the examples in Table 3. The number of tweets that expresses at least one emotion is the 78% of all tweets, which is a fairly high coverage. Also, the number of tweets that express two or more emotions represents the 19% of all tweets.

On top of this, the dataset is annotated with the innovative layer concerning the *target*, including the `topic` (describing the events of the emission) and `direction` (whether messages are directed to a specific

[3]It is important to note that user data is not disclosed, since all data has been anonymized by removing all personal information such as @usernames and generating new IDs for the texts coming from Twitter.

Label	Train	Test (in-domain)	Test (out-of-domain)
Anger	367	56	122
Anticipation	547	85	45
Disgust	874	165	152
Fear	91	13	2
Joy	650	100	98
Love	633	103	135
Neutral	1,322	210	42
Sadness	545	95	132
Surprise	591	102	89
Trust	1,665	272	584
Direction	1,698	260	890
Topic	3,805	671	236
Neither	978	149	9

Table 2
Number of annotations, detailed by label and dataset fold.

Text	Joy	Ang.	Trust	Neu.	Sad.	Love	Dis.	Sur.	Ant.	Fear	Top.	Dir.
Caspita che meraviglia [Gosh what a wonder] #LAmicaGeniale	1	0	1	0	0	0	0	0	0	0	0	0
Queste persone mi spezzano il cuore [These people break my heart] #amorecriminale	0	0	0	0	1	0	0	0	0	0	1	0
il sabato sera con [Saturday evening with] #albertoangela #viaggiosenzaritorno #leggirazziali - Watching Ulisse	0	0	0	0	0	1	0	0	0	0	0	1
Ma i genitori di questi idioti non li hanno mai mandati a scuola? [But didn't the parents of these idiots ever send them to school?] #pechinoexpress	0	1	0	0	0	0	1	0	0	0	1	0

Table 3
Excerpt of the in-domain dataset. The annotations are the following: Text, Joy, Anger, Trust, Neutral, Sadness, Love, Disgust, Surprise, Anticipation, Fear, Topic, Direction.

entity related to RAI) labels. These annotations offer a novel perspective on the data, allowing participants and, in general, the EVALITA community, to explore the effectiveness of current models to understand such a subtask. In total, 84% of the tweets are annotated with the "topic" or "direction" labels, and 8% of tweets have both labels. These annotations should be used for Subtask B.

Dataset for out-of-domain evaluation.
We provide as a second test set 1,000 out-of-domain instances for both subtasks A and B. This additional dataset is composed of comments to music videos and advertisement posted on YouTube. The selection of the videos followed the same procedure used for the creation of the MultiEmotions-It dataset [11]. Specifically, the videos were manually chosen from the songs of Sanremo Music Festival 2021 and from the most recent advertisements, covering different types of products and services. The annotation was performed manually using the same approach of the in-domain dataset. Examples are given in Table 4. In this way, we propose the use of data from a variety of sources that do not directly address RAI contents, but describe other audiovisual media.

As a summary, Table 2 shows the arrangement of the proposed datasets for subtasks A and B, with the detail for each class.

4. Evaluation

In EMit 2023, participants are allowed to submit up to 2 runs for each subtask, with a mandatory run for the main Subtask A. The first run is required to be a constrained submission. That is, the only annotated data to be used for training and tuning the systems are those distributed by the organizers, with the exception of additional data such as lexicons and word embeddings. On the contrary, the second run of each participant can be unconstrained, thus allowing participants to use additional training data. The performance of the systems is evaluated using the **macro-averaged F1-score**, which aggregates the classification metrics for each of the classes thus, in the official

Text	Joy	Ang.	Trust	Neu.	Sad.	Love	Dis.	Sur.	Ant.	Fear	Top.	Dir.
Vergognatevi! Che schifo [Shame on you! How disgusting]	0	1	1	0	0	0	1	0	0	0	1	0
sento la mancanza delle mie crociere .grazie del video e speriamo presto di partire. [I miss my cruises .thanks for the video and hope to go soon.]	0	0	0	0	1	0	0	0	1	0	0	1
Ma quanto è bello Damiano raga? Canzone spaziale ! [But how beautiful is Damiano raga? Space song !]	0	0	1	0	0	0	0	0	0	0	1	1
Adoro questa canzone complimentissimi [I love this song congratulations]	1	0	1	0	0	0	1	0	0	0	1	1

Table 4
Excerpt of the out-of-domain dataset. The annotations are the following: Text, Joy, Anger, Trust, Neutral, Sadness, Love, Disgust, Surprise, Anticipation, Fear, Topic, Direction.

ranking, participants' runs are ordered according to the mentioned F-score.

As baseline, we provide the results of three basic models. All these models compute different text representations that are fed to a logistic regression classifier. In this way, the baselines' text representations are:

- *Baseline_OHE*: uni and bi-grams encoded with a one-hot schema, with a vocabulary of 5,000 tokens.
- *Baseline_TFIDF*: uni and bi-grams represented with the TF-IDF approach, again using a vocabulary of 5,000 tokens.

Finally, we also consider the results of a simple random baseline *Baseline_random*, that outputs the predictions for all classes following a uniform random distribution.

5. Task Overview: Systems and Results

In this first edition of EMit, very few teams participated in the competition. In particular, we received 1 submission by industry (App2Check) and 3 by academic teams (extremITA, ABCD, and EmotionHunters). Although the few participants, the organized shared task also collected international interest with the ABCD team coming from Vietnam. All 4 participating teams have submitted at least one run for Subtask A, and just one team sent us the predictions on Subtask B.

5.1. Systems

Attending to the various systems employed for the classification of emotions (multilabel) and target (binary), their design is based mainly on the use of Large Language Models (LLMs), confirming the actual tendency and success of transformer-based models. However, they have been included in different architectures.

The most used approach is supervised, with a predominance of fine-tuning actions of LLMs to address the specific task of classification. Moreover, two teams also presented semi-supervised systems based specifically on a few-shot prompting (extremITA and App2Check). Various LLMs are employed. For instance, some teams experimented with the classic BERT-based models for the Italian language (i.e., `bert-base-italian-cased`, `bert-base-italian-xxl-cased`, `bert_uncased_L-12_H-768_A-12_italian_alberto`, `umberto-commoncrawl-cased-v1`), others with already fine-tuned versions of BERT (i.e., `feel-it-italian-emotion`, `polibert_sa`), and the rest exploited some sequence-to-sequence LLMs oriented, in this context, to perform mainly instruction solutions such as ChatGPT (`gpt-3.5-turbo-0301`), `flan-t5-xl`, `mt5-base`, IT5 (`it5-efficient-small-el32`) and LLaMA foundational model (`llama-7b-hf`).

In particular, EmotionHunters [12] performed a battery of experiments with classic BERT models and already fine-tuned versions of LLMs. The final system, selected on the basis of their experiments, is based on the fine-tuning of AlBERTo model and, at the top, the fully connected layer to provide a multilabel classification for each text. Both ABCD [13] and App2Check [14] teams employed an ensemble of predictions of different LLMs based on a soft voting method that considers the confidence score associated with each prediction (ABCD: run 1) and the best top-performing model for each emotion (App2Check: unsubmitted run)[4] looking at the performance in the development set of the two best implemented systems: A2C-mT5-r1 (App2Check, run 1) and A2C-GPT-r2 (App2Check, run 2). A2C-mT5-r1 is based on the fine-tuning of multilingual T5 employing the Simple Transformers library. While A2C-GPT-r2 is built using a few-shot approach with ChatGPT, prompt to simultaneously identify all emotions for each text input. A similar approach is used by extremITA [15], who em-

[4]The unsubmitted run reported very good scores in both in-domain (f1-score of 0.504) and out-of-domain setting (f1-score of 0.518)

Ranking	Team name	run id	Anger	Anticipation	Disgust	Fear	Joy	Love	Neutral	Sadness	Surprise	Trust	Macro-average
1	extremITA	2	**0.5176**	**0.6420**	**0.6278**	**0.5833**	0.6178	0.5190	**0.7035**	**0.6258**	**0.5059**	0.6854	**0.6028**
2	extremITA	1	0.4815	0.5594	0.5731	0.1429	0.5909	0.4503	0.6565	0.5233	0.4198	**0.6884**	0.5086
3	ABCD	1	0.4706	0.5946	0.5524	0.0000	**0.6429**	0.4586	0.6462	0.5963	0.3810	0.6516	0.4994
4	EmotionHunters	1	0.4596	0.5205	0.5842	0.2400	0.4589	0.5000	0.4319	0.5484	0.4601	0.6319	0.4835
5	App2Check	2	0.4048	0.3814	0.5831	0.2642	0.3614	**0.5463**	0.3465	0.5181	0.1250	0.2108	0.3741
6	App2Check	1	0.3529	0.4149	0.3855	0.4000	0.6122	0.4867	0.5340	0.4339	0.3293	0.5741	0.4523
7	*baseline_TFIDF*		*0.2945*	*0.4444*	*0.4680*	*0.3684*	*0.3493*	*0.3314*	*0.5392*	*0.3360*	*0.3486*	*0.5944*	*0.4074*
8	*baseline_OHE*		*0.2178*	*0.4221*	*0.3526*	*0.2593*	*0.2918*	*0.3032*	*0.4564*	*0.3191*	*0.2158*	*0.5243*	*0.3362*
9	*baseline_random*		*0.1039*	*0.1541*	*0.2683*	*0.0304*	*0.1760*	*0.1529*	*0.2941*	*0.1565*	*0.2013*	*0.3426*	*0.1872*

Table 5
In-domain evaluation for Task A in terms of f1-score.

Ranking	Team name	run id	Anger	Anticipation	Disgust	Fear	Joy	Love	Neutral	Sadness	Surprise	Trust	Macro-average
1	extremITA	2	0.4051	0.4923	0.6684	0.0000	**0.4416**	**0.7552**	**0.6355**	0.3049	**0.4138**	0.8603	**0.4977**
2	EmotionHunters	1	0.3671	**0.6053**	0.6364	0.0000	0.3768	0.5907	0.6250	0.3103	0.3692	**0.8632**	0.4744
3	extremITA	1	0.5027	0.3667	0.6219	0.0000	0.3176	0.7273	0.5634	0.2024	0.3350	0.8545	0.4491
4	App2Check	1	0.2710	0.4301	0.4691	0.0000	0.4167	0.6528	0.3448	0.2653	0.3662	0.8064	0.4022
5	App2Check	2	**0.6379**	0.3256	**0.6790**	**0.1818**	0.2545	0.6381	0.2564	**0.3195**	0.1373	0.3001	0.3730
6	*Baseline_OHE*		*0.3972*	*0.4533*	*0.4197*	*0.0000*	*0.1890*	*0.3218*	*0.1974*	*0.3129*	*0.2812*	*0.7468*	*0.3319*
7	*Baseline_TFIDF*		*0.3342*	*0.4412*	*0.4092*	*0.0000*	*0.1786*	*0.3034*	*0.1778*	*0.2649*	*0.1913*	*0.6869*	*0.2987*
8	*Baseline_random*		*0.1984*	*0.0917*	*0.2188*	*0.0081*	*0.1624*	*0.2208*	*0.0841*	*0.2097*	*0.1336*	*0.5483*	*0.1876*

Table 6
Out-of-domain evaluation for Task A in terms of f1-score.

ployed sequence-to-sequence LLMs for Italian to solve instructions related to specific tasks. They developed two systems to solve different shared tasks of EVALITA 2023: extremIT5 (extremITA, run 1) and extremITLLaMA (extremITA, run 2). The former is an Encoder-Decoder model based on IT5, and trained by concatenating the task name and an example as input (i.e., "EMit: Quando ci sarà l'espulsione di Claudia #ilcollegio [url]") and as output the sequence of labels; in contrast, the latter is an instruction-tuned Decoder model built upon the LLaMA foundational models, therefore the structured prompt is an instruction in natural language like "Which emotions are expressed in this text? You can choose among joy, fear, ...". Differently from the previous editions of EVALITA, in the EMit 2023 shared task it is clear that the attention is only on the LLMs' ability to solve tasks and their integration into the systems' architecture, losing the focus on linguistic features that can represent or infer the emotions in the text. Also, the preprocessing of the text is focused on very few steps, regarding mainly the transformation of emojis in textual descriptions, removing mentions, urls and other symbols.

5.2. Results

Tables 5, 6, 7, and 8 report the official results obtained in EMit 2023 for both subtask A and B. The ranking is based on the macro-averaged F1-score, and considers both the team and the run of each submission. The higher scores for each column are marked in bold, while the lower scores are underlined.

Generally, it is interesting to see that even if the classification problem of Subtasks A and B are very different, the best results for each are similar. Concretely, when considering the in-domain test set, the best submission for Subtask A obtained a macro-averaged score of 0.6028, while for Subtask B it is 0.6459. In the case of the out-of-domain evaluation, the best scored obtained by a team is 0.4977 in Subtask A, and 0.4448 in Subtask B. This decrease in the classification performance when comparing in-domain and out-of-domain evaluations was expected giving that training was performed only on the in-domain data. Additionally, it is worth noticing that even if Subtask A contains 10 possible labels and Subtask B has only 2, their best scores are not that different (a difference of 0.0431).

Following, in relation to the overall results achieved by participants, it can be seen that in Subtask A, both in the in-domain and out-of-domain evaluations, the teams' submissions have obtained better results than the baselines. The best baseline in the in-domain evaluation uses TF-IDF uni and bi-grams, while for the out-of-domain evaluation the uni and bi-grams using one-hot encoding achieves the best result. Regarding Subtask B, the only team that has submitted a run for it has obtained a better score than in the in-domain evaluation.

In contrast, when considering the out-of-domain evaluation in Subtask B, we see that the best baseline is the one that randomly predicts the objective labels. This decreases in the classification performance is seen in the runs but also in the learning-based baselines. This may be explained by considering the distribution of the out-of-domain sets in Subtask B (see Table 2). Indeed, we can observe that in the train and in-domain test sets the prevalent label is `Topic` but, conversely, in the out-of-domain test set the `Direction` label is more frequent. Consequently, it is possible to postulate that systems trained with the Subtask B training set would perform fairly well in the in-domain test set, but worse on the out-of-domain data.

Ranking	Team name	run id	Direction	Topic	Macro-average
1	extremITA	2	0.4855	**0.8064**	**0.6459**
2	extremITA	1	0.4963	0.7699	0.6331
3	*Baseline_TFIDF*		***0.5032***	*0.7336*	*0.6184*
4	*Baseline_OHE*		*0.4651*	*0.7212*	*0.5932*
5	*Baseline_random*		*0.3443*	*0.5614*	*0.4528*

Table 7
In-domain evaluation for Task B in terms of f1-score.

Ranking	Team name	run id	Direction	Topic	Macro-average
1	*Baseline_Random*		*0.6452*	***0.3651***	***0.5051***
2	*Baseline_TFIDF*		*0.5559*	*0.3573*	*0.4566*
3	extremITA	1	**0.6831**	0.2066	0.4448
4	*Baseline_OHE*		*0.4884*	*0.3571*	*0.4228*
5	extremITA	2	0.3275	0.3172	0.3223

Table 8
Out-of-domain evaluation for Task B in terms of f1-score.

Finally, the detailed results of the evaluation offer interesting insights into the models' performance. For example, when considering the effect of the number of instances for each class (Table 2, we see that in Subtask A `Fear` is much less frequent in comparison to the other emotions. Hence, this has an effect on the performance of the systems: in the out-of-domain evaluation (Table 6) the majority of the models obtained a null score in the `Fear` category, thus affecting in a negative way the overall averaged score. Similarly, the most common emotions in Subtask A (`Trust` and `Neutral`) are generally better predicted by the participants' systems.

6. Discussion

The presence of both in-domain and out-of-domain data in the EMit task provides a valuable experimentation setting as proved by the different performances in classification between the two evaluations settings. Since these two types of datasets have been obtained from different sources (see Sect. 3), they represent a diverse collection of cases. In this way, we can evaluate the participants' models in relation to their generalization capabilities.

In fact, we observe a general reduction in the classification metrics when comparing the in-domain and out-of-domain test sets. In Subtask A, with the in-domain set, the average macro f-score of all participants' systems is 0.4868. In comparison, the average metric drops to 0.4393 in the case of the out-of-domain dataset. We can see a similar trend when considering Subtask B, even if just one team has participated. The average score in the in-domain evaluation is 0.6395 and, in the out-of-domain case, 0.3935.

While participants have achieved promising results in the detection of emotions and opinion targets, there is still room for improvement. The large number of emotions considered in Subtask A is indeed a challenge for automatic systems, increasing the difficulty of the task. In comparison, Subtask B has fewer categories, but still, the proposed systems and baselines obtain rather low metrics in the task. Also, we have seen how the representation of the different emotions greatly impacts classification performance. These, along with the generalization difficulties in the out-of-domain set, indicate that the challenge proposed in EMit is not solved. Indeed, future works need to address the shortcomings detected and advance in the generation of systems that are more robust to the frequency of categories in the datasets, as well as the inclusion of domain-specific knowledge that may improve overall results.

7. Conclusions

The first edition of EMit (Emotions in Italian) proposes the assessment of emotions on Italian texts by presenting an interesting challenge that revolves around two subtasks. On one hand, the main task (subtask A) presents a comprehensive emotion annotation set using Plutchik's model, with the addition of the *love* emotion. On the other hand, subtask B introduces a novel classification problem, which addresses the target of the opinion expressed in the textual message. To complement this, we also provide out-of-domain test sets to further obtain insights into the behaviour of the participants' systems.

To advance in the study of opinion mining in relation to emotion, and considering both subtasks, EMit establishes a rich annotation schema for considering the effect of this challenge on automated systems. While only one team participated in subtask B, we believe that the additional perspectives brought by the combined study of emotions and their targets will be the subject of further

studies. As an example, an interesting research avenue could study the variation of emotions depending on the target, and how this affects learning systems. Another potential research direction is the inclusion of linguistic knowledge into the commonly used large language models.

Acknowledgments

The work of Oscar Araque has been partially funded by the Spanish Ministry of Science, Innovation, and Universities through project COGNOS (PID2019-105484RB-I00) and "ETSI Telecomunicación" of "Universidad Politécnica de Madrid" through the initiative "Primeros Proyectos" under "AFRICA – Detecting and Analyzing Affective and Moral Factors in Radicalization and ExtremIsm: a MaChine learning Approach". The work of S. Frenda and V. Patti was partially funded by the Multilingual Perspective-Aware NLU Project in partnership with Amazon Alexa. The work of D. Nozza was partially funded by Fondazione Cariplo (grant No. 2020-4288, MONICA).

References

[1] A. Origlia, V. Galatà, Evalita 2014: Emotion recognition task (ERT), in: Proceedings of the Fourth International Workshop EVALITA 2014, Pisa University Press, 2014, pp. 112–115.

[2] 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.

[3] 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.

[4] M. G. Vega, M. C. Díaz-Galiano, M. Á. G. Cumbreras, F. M. P. del Arco, A. Montejo-Ráez, S. M. J. Zafra, E. M. Cámara, C. A. Aguilar, M. A. S. Cabezudo, L. Chiruzzo, et al., Overview of TASS 2020: Introducing emotion detection, in: IberLEF@ SEPLN, 2020.

[5] F. M. Plaza-del Arco, S. M. Jiménez Zafra, A. Montejo Ráez, M. D. Molina González, L. A. Ureña López, M. T. Martín Valdivia, Overview of the emoevales task on emotion detection for spanish at iberlef 2021 (2021).

[6] 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.

[7] C.-C. Hsu, L.-W. Ku, SocialNLP 2018 EmotionX challenge overview: Recognizing emotions in dialogues, in: Proceedings of the sixth international workshop on natural language processing for social media, 2018, pp. 27–31.

[8] V. Barriere, S. Tafreshi, J. Sedoc, S. Alqahtani, WASSA 2022 shared task: Predicting empathy, emotion and personality in reaction to news stories, in: Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 214–227. URL: https://aclanthology.org/2022.wassa-1.20. doi:10.18653/v1/2022.wassa-1.20.

[9] 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.

[10] R. Plutchik, H. Kellerman, Theories of emotion, volume 1, Academic Press, 1980. URL: https://books.google.it/books?id=TV99AAAAMAAJ.

[11] 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.

[12] G. Calò, F. Massafra, B. De Carolis, C. Loglisci, TASK A at EVALITA 2023: Overview of the Emotion Hunters Approach to the Categorical Emotion Detection in Italian Social Media, 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.

[13] D. Nguyen Ba, U. N. Ngoc Phuong, T. Dang Van, Ensemble Approach for Categorial Emotion Detection in Social Media Messages: EMit at EVALITA 2023, 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.

[14] G. Cageggi, E. Di Rosa, A. Uboldi, Large Language Models for Multilabel Emotion Classification: fine-tuned LLM vs plain FLAN and ChatGPT, in: Proceedings of the Eighth Evaluation Campaign of Nat-

ural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[15] C. D. Hromei, D. Croce, V. Basile, R. Basili, ExtremITA at EVALITA2023: Multi-Task Sustainable Scaling to Large Language Models at its Extreme, 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.