EMit 2023: Categorical Emotion Detection in Italian

Oscar Araque
Universidad Politécnica de Madrid
Madrid, Spain

Simona Frenda
Università degli Studi di Torino
Turin, Italy

Debora Nozza
Università Bocconi
Milan, Italy

Viviana Patti
Università degli Studi di Torino
Turin, Italy

Rachele Sprugnoli
Università di Parma
Parma, Italy

May 8, 2023

Contents

1 Introduction

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 [7]. 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) [10] 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 [6] 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 [11] and EmoEvalEs at IberLEF 2021 were only on Spanish tweets [1]; instead, *EmoContext* at SemEval 2019 [3] and *EmotionX* at the SocialNLP workshop in 2018[5] 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 [1].

In this context, the **EMit (Emotions in Italian)** task aims at providing the first evaluation framework for emotion detection in Italian texts and make new annotated data available to the community.

2 Task Description

EMit is organized according to two subtasks:

- **Subtask A**
  Emotion Detection (Main Task) The main proposed subtask is the detection of emotions in social media messages about TV shows, series 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 should be 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*, it means that the text is about events, issues discussed in the television episode/music video/advertisements, or invited guests of a tv show. When the target is the *direction*, this implies that the message is about 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**, thus, the participating system should provide in the output the *id* of the message and all the possible labels contained in it (see Section 4). In Subtask A, the message could be classified as neutral, or expressing one or more emotions. In Subtask B, the message could be classified as addressing the topic, or the direction, or both or neither. For Subtask A, the provided labels are: "neutral" when the message does not express emotions, the 8 main emotions defined by Plutchik [8] (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. Therefore, a total of 10 labels should be used for Subtask A. For Subtask B the provided labels are: "topic" and "direction".

Considering the specific attention on the entertainment sector, we designed the Subtask B specific 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 to develop a sort of *Auditel of emotions*).

In order to evaluate the robustness of the models proposed by participants, we will release two different test sets, one (*in-domain*) including tweets of the same textual genre and subjects of the training set, and another one including social data of different genres and subjects (*out-of-domain*), offering to participants a cross-domain evaluation setting for both subtasks A and B.

3 Dataset

We will exploit two datasets that are intended to evaluate different aspects of the participants' models, evaluated according to an *in-domain* and *out-of-domain* setting.

Dataset for in-domain evaluation.
This dataset is obtained from Twitter, and it is composed of approximately 6,966 tweets that discuss programs by the Italian RAI TV station. Such messages have been annotated by three independent annotators with a multi-layered annotation scheme. Concerning the emotion layer, each tweet has been annotated according to the 8 basic emotions defined by Plutchik [8] (anger, anticipation, disgust, fear, joy, sadness, surprise, trust), and the additional label "love". To indicate the absence of emotion, the "neutral" label is used. These emotion annotations should be used for running Subtask A.

The emotion annotations are non-exclusive, thus a certain tweet can be annotated with one or more emotions, or even solely the neutral label (see examples in Table 3). The number of tweets that express at least one emotion is the 78% of all tweets, which is a fairly high coverage. 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 the message is directed to a specific 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 tweet are annotated with the "topic" or "direction" labels. These annotations should be used for Subtask B.

Dataset for out-of-domain evaluation.
This dataset will be composed of messages that are not extracted from the same origin, but are related to the content of the in-domain dataset. 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. We will provide as a second test set, out-of-domain instances for both subtasks A and B.

Table 1 shows the arrangement of the proposed datasets for subtasks A and B.

Learning Set	Dataset	Total (approx.)
Subtask A		
Train	In-domain	5,966
Test 1	In-domain	1000
Test 2	Out-of-domain	1000
Subtask B		
Train	In-domain	5,966
Test 1	In-domain	1000
Test 2	Out-of-domain	1000

Table 1: EMit datasets and their distribution in tasks and learning sets.

text	Joy	Anger	Trust	Neutral	Sadness	Love	Disgust	Surprise	Anticipation	Fear	Topic	Direction
"Caspita che meraviglia #LAmicaGeniale"	1	0	1	0	0	0	0	0	0	0	0	0
"Queste persone mi spezzano il cuore 💔 #amorecriminale"	0	0	0	0	1	0	0	0	0	0	1	0
"Il sabato sera con #albertoangela ❤️❤️ #viaggiosenzaritorno #leggirazziali - Guardando 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? #pechinoexpress"	0	1	0	0	0	0	1	0	0	0	1	0

Table 2: Excerpt of the proposed in-domain dataset.

Participants will be allowed using other few existing and available Italian datasets annotated for emotions, that could be used as additional training data in the *unconstrained* runs:

- MultiEmotions-It [9];

- FEEL-IT [2];

- AriEmozione 2.0 [4].

3.1 Release

Given the restrictions posed by the EU GDPR regulation both training and test data are made available **for academic purposes only**, therefore participants will access the data by filling in the online form published in the task website, and by accepting the task's terms and conditions. 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.

4 Submission of Results

As explained before, both subtasks are designed as **multilabel problems of classification**, thus:

- in Subtask A, the message could be classified as neutral, or expressing one or more emotions;

- in Subtask B, the message could be classified as addressing the topic, or the direction, or both or neither.

4.1 Submission format

The participants should provide an output file in `csv` format reporting:

- the name of the team,

- the addressed Subtask (`A` or `B`),

- the ID of the run (`1` or `2`)[2],

- and the genre of test set (`in-domain` or `out-of-domain`).

For instance, if in the first run the team addresses the Subtask A on the in-domain setting, the name of the file should be:

<div align="center">

`TeamName_SubtaskA_in-domain_run1.csv`

</div>

If in the first run, the team addresses the Subtask A on the out-of-domain setting, the name of the file should be:

<div align="center">

`TeamName_SubtaskA_out-of-domain_run1.csv`

</div>

The examples for the submission files for both subtasks are reported in the EMit repository on GitHub: https://github.com/oaraque/emit.
These files contain only the **id** of the message and all the labels of each subtask (see Figures 1 and 2).

The list of labels for Subtask A contains: 'Anger', 'Anticipation', 'Disgust', 'Fear', 'Joy', 'Love','Neutral', 'Sadness', 'Surprise', and 'Trust'.

id	Anger	Anticipation	Disgust	Fear	Joy	Love	Neutral	Sadness	Surprise	Trust
246b5a294b3208be6800067b9f0f9e87	1	0	1	0	0	1	0	0	0	0
611f78bf3db6c71f0ed053459889fd4b	0	0	1	0	1	1	1	1	1	0
d37b6e59c8ba22a23755566c7797f69a	0	1	0	1	0	1	1	1	0	1
60cbfcc1eeaed1d8adf8fd7b3bc17dba	1	0	0	1	1	1	0	0	1	0

Figure 1: Example of the csv format for Subtask A

The list of labels for Subtask B contains: 'Direction' and 'Topic'.

id	Direction	Topic
246b5a294b3208be6800067b9f0f9e87	1	0
611f78bf3db6c71f0ed053459889fd4b	1	0
d37b6e59c8ba22a23755566c7797f69a	0	0
60cbfcc1eeaed1d8adf8fd7b3bc17dba	0	1

Figure 2: Example of the csv format for Subtask B

Important: all the relevant runs must be compressed as a single ZIP file named:

<div align="center">

`TeamName.zip`

</div>

4.2 How to submit your runs

The ZIP file containing the relevant runs should be submitted during the evaluation window (2nd-19th May 2023) to the following email address:

<div align="center">

emit.evalita2023@gmail.com

</div>

---

[2]Participating teams will explain which runs are constrained or unconstrained in their final reports.

with the subject:

<center>"Submission: EMit - EVALITA2023 - TeamName"</center>

It is important to notice that:

- Participation is allowed to all two subtasks (Subtask A and Subtask B) or only to Subtask A.

- The participating teams should submit the predictions on both test sets (*in-domain* and *out-of-domain*), regardless if they participate to Subtask A or both subtasks.

- Each participating team will be allowed to submit a **maximum of 2 runs** for each subtask. **Submitting at least one run on subtask A will be anyway compulsory**.

- The first run will be produced according to the *constrained* modality: the only annotated data to be used for training and tuning the systems are those distributed by the organizers. Other linguistic resources, e.g., word embeddings, and lexicons are instead allowed.
  The second run could be produced according to the *unconstrained* modality: annotated external data can be also employed. All external resources should be described in the systems' reports.
  The constrained run is compulsory, while the unconstrained run is optional.

5 Evaluation

During the evaluation window (2nd-19th May 2023), the unlabelled test data will be released; and only after the assessment, the gold labels will be released as well.

For each task, a separate official ranking will be provided and the evaluation will be performed using the established measures for such tasks, namely, Precision, Recall, and F1 score per class [6].

The ranking of the systems will be done using **macro-averaged F1 scores** for both subtasks.

Finally, the baselines' systems for both subtasks have been built using a logistic regression with stochastic gradient descent learning that take as input a representation of texts based on unigrams and bigrams of words weighted with TF-IDF (see the repository on GitHub).

6 Important Links

- Webpage of the EMit shared task:
  http://www.di.unito.it/~tutreeb/emit23/index.html

- The training set, the examples for the submission files, and the evaluation script are available on the repository of GitHub:
  https://github.com/oaraque/emit

- Form for obtaining the password for the training set:
  https://forms.gle/6tHUZ4qcwLfkNior9

- For any question or problem, please open a thread on the Google Group at
  emit_evalita2023@googlegroups.com

7 Contacts

- Oscar Araque: `o.araque@upm.es`
- Simona Frenda: `simona.frenda@unito.it`
- Rachele Sprugnoli: `rachele.sprugnoli@unipr.it`
- Debora Nozza: `debora.nozza@unibocconi.it`
- Viviana Patti: `viviana.patti@unito.it`

References

[1] Valentin Barriere, Shabnam Tafreshi, João Sedoc, and Sawsan 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, pages 214–227, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[2] Federico Bianchi, Debora Nozza, and Dirk Hovy. FEEL-IT: Emotion and sentiment classification for the Italian language. In Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 76–83, Online, April 2021. Association for Computational Linguistics.

[3] Ankush Chatterjee, Kedhar Nath Narahari, Meghana Joshi, and Puneet Agrawal. SemEval-2019 task 3: EmoContext contextual emotion detection in text. In Proceedings of the 13th International Workshop on Semantic Evaluation, pages 39–48, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics.

[4] Francesco Fernicola, Shibingfeng Zhang, Federico Garcea, Paolo Bonora, and Alberto Barrón-Cedeno. Ariemozione: Identifying emotions in opera verses. In Proceedings of CLiC-it. CEUR-WS, 2020.

[5] Chao-Chun Hsu and Lun-Wei Ku. SocialNLP 2018 EmotionX challenge overview: Recognizing emotions in dialogues. In Proceedings of the sixth international workshop on natural language processing for social media, pages 27–31, 2018.

[6] Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. SemEval-2018 task 1: Affect in tweets. In Proceedings of the 12th International Workshop on Semantic Evaluation, pages 1–17, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[7] Antonio Origlia and Vincenzo Galatà. Evalita 2014: Emotion recognition task (ERT). In Proceedings of the Fourth International Workshop EVALITA 2014, pages 112–115. Pisa University Press, 2014.

[8] Robert Plutchik and Henry Kellerman. Theories of emotion, volume 1. Academic Press, 1980.

[9] Rachele Sprugnoli. Multiemotions-it: A new dataset for opinion polarity and emotion analysis for italian. In 7th Italian Conference on Computational Linguistics, CLiC-it 2020, pages 402–408. Accademia University Press, 2020.

[10] Carlo Strapparava and Rada Mihalcea. SemEval-2007 task 14: Affective text. In Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), pages 70–74, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[11] Manuel García Vega, Manuel Carlos Díaz-Galiano, Miguel Ángel García Cumbreras, Flor Miriam Plaza del Arco, Arturo Montejo-Ráez, Salud María Jiménez Zafra, Eugenio Martínez Cámara, César Antonio Aguilar, Marco Antonio Sobrevilla Cabezudo, Luis Chiruzzo, et al. Overview of TASS 2020: Introducing emotion detection. In IberLEF@ SEPLN, 2020.