FEEL-IT: Emotion and Sentiment Classification for the Italian Language

Federico Bianchi[*]
Bocconi University
Via Sarfatti 25, 20136
Milan, Italy
f.bianchi@unibocconi.it

Debora Nozza[*]
Bocconi University
Via Sarfatti 25, 20136
Milan, Italy
debora.nozza@unibocconi.it

Dirk Hovy
Bocconi University
Via Sarfatti 25, 20136
Milan, Italy
dirk.hovy@unibocconi.it

Abstract

Sentiment analysis is a common task to understand people's reactions online. Still, we often need more nuanced information: is the post negative because the user is angry or because they are sad? An abundance of approaches has been introduced for tackling both tasks. However, at least for Italian, they all treat only one of the tasks at a time. We introduce FEEL-IT, a novel benchmark corpus of Italian Twitter posts annotated with four basic emotions: *anger*, *fear*, *joy*, *sadness*. By collapsing them, we can also do sentiment analysis. We evaluate our corpus on benchmark datasets for both emotion and sentiment classification, obtaining competitive results. We release an open-source Python library, so researchers can use a model trained on FEEL-IT for inferring both sentiments and emotions from Italian text.

1 Introduction

Emotions shape our lives and the way we communicate. We can be happy, sad, or angry, and we can let others know of our emotional state through language. Thus, efficiently detecting emotion in text is essential for analyzing people's position towards a topic. Product and service companies frequently use emotion and sentiment data to inform advertising campaigns and measure customer satisfaction (Ahmad et al., 2020). Emotions have a central role in a political campaigns, and political discourse in particular (Huguet Cabot et al., 2020). Emotion and sentiment recognition can also aid in the critical decision-making process of crisis management or emergency scenarios (Stowe et al., 2016; Desai et al., 2020).

anger	*fear*	*joy*	*sadness*	Total
912	103	728	294	2037

Table 1: FEEL-IT corpus statistics.

Despite the huge interest of the Natural Language Processing community, the majority of benchmark datasets have been proposed for English (Calefato et al., 2017; Abdul-Mageed and Ungar, 2017; Akhtar et al., 2019, inter alia) showing a limited interest for other languages, such as German (Troiano et al., 2019), Chinese (Wang et al., 2018), Spanish (Navas-Loro and Rodríguez-Doncel, 2019), Italian (Barbieri et al., 2016; Sprugnoli, 2020), and multiple languages in shared tasks (Mohammad et al., 2018; Pontiki et al., 2016).a Moreover, they are usually collected either via hashtags and emojis for distant supervision (Abdul-Mageed and Ungar, 2017; Mohammad, 2012; Pak and Paroubek, 2010; Lamprinidis et al., 2021), or via very specific topics (Khanpour and Caragea, 2018; Chang et al., 2018; Nozza et al., 2017). The first causes noisy training data (Bing et al., 2015), the second results in highly domain-specific datasets.

This paper presents FEEL-IT, a novel benchmark corpus of Italian Twitter posts annotated with four basic emotions (Ekman, 1992): *anger*, *fear*, *joy*, *sadness*.[1] To the best of our knowledge, no other Italian dataset with a broad topic and domain coverage for emotion and sentiment classification exists. Beyond releasing benchmark results on FEEL-IT, we evaluate recent neural models trained on our corpus for emotion recognition

[*] Both authors contributed equally to this research and are ordered alphabetically.

[1] We focus on these emotions because they appear most frequently in text.

Proceedings of the 11th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 76–83
April 19, 2021. ©2021 Association for Computational Linguistics

Example	Emotion
Pagliacci ammaestrati dal Grillo parlante di Pinocchio *They are buffoons controlled by Pinocchio's Jiminy Cricket*	*anger*
Non ci sto dormendo la notte. #22Agosto #COVID19 *This does not make me sleep at night. #22August #COVID19*	*fear*
Adoro questa canzone, è una delle mie preferite STREAM ICARUS FALLS *I love this song, it's one of my favourite STREAM ICARUS FALLS*	*joy*
I brividi. Come si può spegnere una vita con così tanta facilità? Non ho parole.... *I got chills. How can you kill someone so easily? I do not know what to say....*	*sadness*

Table 2: Examples of FEEL-IT annotations. English translations are reported in italic.

on the MultiEmotions-It dataset (Sprugnoli, 2020). It contains comments on music videos and advertisements posted on YouTube and Facebook. We also test performance on sentiment classification by collapsing positive and negative emotions on the SENTIPOLC16 benchmark dataset (Barbieri et al., 2016). It comprises both general and political topics. The best-performing models are released as part of a Python library to foster and facilitate research on the topic.

Contributions. We present FEEL-IT, a new corpus on Italian tweets, annotated with four basic emotions (*anger*, *fear*, *joy*, *sadness*). We demonstrate that we can effectively predict sentiments and emotions in text by training prediction models on this corpus. We release an open-source Python library[2] that researchers can use to classify their text.

2 Data Collection and Annotation

We retrieved the data by monitoring trending topics each day between 20th August to 12th October 2020, using the Twitter API. For each day, we sampled 1000 tweets. This approach allowed us to get data from a range of different topics that span over many weeks.

The two first authors labeled the complete set of posts. Both are native Italian speakers with a strong NLP background. Eventually, the number of annotated tweets that contained an emotion was 2037 tweets (we removed tweets that did not contain any emotion, that is, most of them). This process involves a lot of data that has been discarded, and it is time-consuming, but the upside of it is that the collected tweets are from diverse domains and are high quality.

We computed our inter-rater agreement on a shared set of 220 tweets, annotated both with emotions and with *none* (i.e., no emotion found). We reached an agreement of 0.6 (Krippendorff's Alpha). Once *none* tweets were removed, the agreement on the remaining 68 annotated tweets was 0.8 (Krippendorff's Alpha).

Corpus Analysis Table 1 shows the label distribution of the FEEL-IT corpus for the four basic emotions considered. Examples for each class are shown in Table 2.

Similar to other realistic emotion classification datasets (Sprugnoli, 2020; Mohammad et al., 2018; Nozza et al., 2017; Mohammad, 2012), the dataset is imbalanced. The distribution is similar to the SemEval-2018 Task 1 dataset (Mohammad et al., 2018), where anger and joy account for the majority of tweets, and fear is the least frequent emotion.[3]

In FEEL-IT, topics vary both with respect to domains and time. Topic domains ranges from health (*#covid19*, *#mascherina/mask*) to sports (*#F1*, *#Juventus*), from social issues (*#scuola/school*) to TV shows (*#GFvip*, *#pomeriggio5*), from individuals (*#DiMaio*, *#Suarez*) to generic targets (*#negazionisti/negationists*). Each topic is associated with a time range that greatly varies with subject. TV shows are cited when they are broadcast, e.g., *#domenicalive*, literally *Sunday live* is mainly commented on Sunday. Some events, like soccer matches or celebrity birthdays, are mentioned only one day, e.g., the hashtag of the soccer match *#BeneventoInter* appears 371 times, but only the 31st September. Tweets related to COVID-19 are present every observed day, with some peaks for specific events (e.g., on the 2nd October, we recorded a peak of 132 tweets due to the news of US president Trump testing positive for COVID-19).

3 Experiments

We use experimental evaluation to (i) show that our classifier can predict emotions in tweets and (ii)

[2] https://github.com/MilaNLProc/feel-it

[3] Note that in other datasets, joy is the most frequent emotion, because of their focus on music or movies.

that FEEL-IT can also be used to perform sentiment classification with competitive results.

3.1 Emotion Classification

We first experiment with emotion recognition in the FEEL-IT dataset. Contextualized representations, such as BERT (Devlin et al., 2019) have obtained a lot of attention due to the great results (Rogers et al., 2021; Nozza et al., 2020) on multiple languages and on different tasks (Scarlini et al., 2020; Mass and Roitman, 2020; Du et al., 2020; Pasini et al., 2020; Peinelt et al., 2020; Bianchi et al., 2021; Nozza et al., 2021, inter alia). In this paper, we use the Italian BERT model UmBERTo trained on Commoncrawl ITA.[4] As the first experimental condition, we fine-tune the UmBERTo model for the task of emotion classification with the considered training data (*UmBERTo-FT*).

As additional experimental frameworks, we use three different approaches to represent tweets: (i) We collect pre-trained UmBERTo representations using average pooling of the last layer (*UmBERTo-PT*); (ii) we use an Italian word2vec model (*W2V*)[5] and create the representation of the tweet as the average of the word embeddings; (iii) we use a *TF-IDF* baseline with bi-grams to represent tweets. To make TF-IDF and W2V as competitive as possible, we apply a pre-processing pipeline to the text: (1) replace URLs and mentions with unique tokens; (2) replace emojis with a description of the emoji (Leonardelli et al., 2020), (3) split hashtags on camel case (#HappyBirthday becomes Happy Birthday); (4) remove punctuation. Given the representations, we use logistic regression with a softmax and with instance-weight balancing for classification. We test the models in a 10-fold cross-validation setting. We use the Most Frequent Class (*MFC*) as the baseline method.

Results. Table 3 reports Precision, Recall, F1-score, and Accuracy of the different tested models. First, we see that all the proposed models overcome the MFC baseline. Second, we observe that UmBERTo-FT is the model that obtains the best results in terms of the overall performance metrics.

We can draw further insights from the class-wise F1-score shown in Table 4. As expected, the emotion classes with the most training instances (*joy* and *fear*) are also the ones on which the classifiers perform best. Again, UmBERTo-FT is the model with the highest overall performance.

Model	P	R	F1	Acc
UmBERTo-FT	**0.72**	**0.73**	**0.71**	**0.82**
UmBERTo-PT	0.64	0.67	0.65	0.76
W2V	0.57	0.62	0.58	0.76
TF-IDF	**0.72**	0.60	0.64	0.74
MFC	0.11	0.25	0.15	0.45

Table 3: Precision (P), Recall (R), F1-score (F1), and Accuracy (Acc) of the cross-validated emotion classification on FEEL-IT.

Model	*anger*	*fear*	*joy*	*sadness*
UmBERTo-FT	**0.88**	**0.51**	**0.87**	**0.60**
UmBERTo-PT	0.85	0.38	0.86	0.51
W2V	0.80	0.31	0.76	0.42
TF-IDF	0.80	**0.51**	0.80	0.46

Table 4: F1-score per class of the cross-validated emotion classification on FEEL-IT.

The only emotion for which UmBERTo-FT obtains lower equal to TF-IDF is *fear*. It should be noted that this is the least frequent class in the dataset and, therefore, the more difficult to capture. The different prediction behavior on this class is also why the large difference in precision in Table 3. Indeed, precision for the class *fear* is 0.76 for TF-IDF, 0.55 for UmBERTo-FT, and 0.33 for UmBERTo-PT, while recall is 0.38, 0.52, and 0.53, respectively. This discrepancy means that, while TF-IDF is more cautious on assigning the label *fear*, UmBERTo-FT and UmBERTo-PT have a high number of false positives (see Appendix B for confusion matrices). From a qualitative perspective, we see that many of these false-positive tweets could be associated with *fear*, even if the most prevalent emotion is *anger* or *sadness*. This correspondence indicates that tweet authors tend to communicate their fears by other, less intimate, emotions. Examples are "Siete un branco di egoisti che pensa solo al proprio, fregandosene di mettere a rischio la vita di tutti gli altri" (*You are a bunch of selfish people who only think about themselves, not caring about putting everyone else's life at risk*) and "Ogni giorno compilo il mio excel sulla situazione in Veneto...e ogni giorno lo chiudo pensando Speriamo che domani ci siano dati un po' più incoraggianti" (*Every day I fill an excel file on the situation in Veneto...and every day I close it thinking "Let's hope that tomorrow we are going to have more encouraging data"*).

[4] https://github.com/musixmatchresearch/umberto

[5] http://vectors.nlpl.eu/repository/20/52.zip, see (Fares et al., 2017)

	Training data	P	R	F1	Acc
FT	SP16	0.79	**0.84**	0.80	0.82
	FEEL-IT	**0.82**	0.80	**0.81**	**0.84**
	SP16+FEEL-IT	0.80	**0.84**	**0.81**	0.82
PT	SP16	0.77	0.82	0.77	0.77
	FEEL-IT	0.81	0.80	0.80*	**0.84**
	SP16+FEEL-IT	0.78	0.83	0.79	0.80

Table 5: Precision (P), Recall (R), F1-score (F1), and Accuracy (Acc) of sentiment classification on SENTIPOLC16 using UmBERTo-FT and UmBERTo-PT model. We tested the statistical significance of the *F1-score for UmBERTo-PT trained on FEEL-IT showing that it is significantly better than the one trained on SP16 (bootstrap sampling $p < 0.05$).

3.2 Sentiment Analysis

We test on SENTIPOLC16 (SP16) (Barbieri et al., 2016) to evaluate the performance of sentiment classification models trained on FEEL-IT. We collapsed the FEEL-IT classes into 2 by mapping *joy* to the *positive* class and *anger*, *fear* and *sadness* into the *negative* class. We use the fine-tuned UmBERTo model (*UmBERTo-FT*) and the logistic regression classifier applied to its representations (*UmBERTo-PT*).

SP16 also comes with a training set. We fit a classifier on this data to see whether it is better to train on FEEL-IT or SP16. Eventually, we also combine the two datasets to see if we can get the best of both worlds. SP16 comprises tweets that could be both positive and negative; in our experiment, we exclude the tweets that were labeled both positive and negative. Thus, SP16 training contains 4154 examples, while the test contains 1050 samples.

Results. Table 5 shows the results for the sentiment classification task. They demonstrate that our proposed corpus is useful for sentiment prediction. While FEEL-IT contains roughly half of the tweets that SP16 has, the performance obtained with FEEL-IT on the SP16 test set is the best. Interestingly, using a more sophisticated model (UmBERTo-FT) leads to narrowing the differences between performance. This result confirms that our dataset can 1) be used for sentiment analysis and 2) obtains state-of-the-art performances on the current benchmark for Italian sentiment analysis. Note that the combination between SP16 and FEEL-IT brings good recall, with a slight drop in Precision and F1-score.

Model	Testing	P	R	F1	Acc
UmBERTo-FT	ME	**0.56**	0.59	**0.57**	**0.73**
UmBERTo-PT	ME	**0.56**	**0.66**	**0.57**	0.69
MFC	ME	0.16	0.25	0.20	0.64
UmBERTo-FT	C19	**0.56**	0.56	**0.56**	**0.69**
UmBERTo-PT	C19	0.53	0.53	0.50	0.60
MFC	C19	0.15	0.25	0.19	0.60

Table 6: Precision (P), Recall (R), F1-score (F1), and Accuracy (Acc) of emotion recognition on use-cases using UmBERTo model.

4 Use-cases: COVID-19 and MultiEmotions-It

To further validate our approach, we showcase the results of emotion recognition models trained on FEEL-IT and tested on two topic-specific datasets: MultiEmotions-It (Sprugnoli, 2020) and a dataset of 662 tweets about COVID-19.

MultiEmotions-It (ME) (Sprugnoli, 2020) is a linguistic resource for Italian which comprises comments of music videos and advertisements posted on YouTube and Facebook. Each text is manually annotated according to four different dimensions: i.e., relatedness, opinion polarity, emotions, and sarcasm. This dataset differs from FEEL-IT both in terms of topic variety and considered social media. Among all the emotion classes considered in ME, we removed the ones not pertaining to our set of emotions. After this process, we are left with 304 comments.

As before, we pick UmBERTo-FT and UmBERTo-PT as our champion models. To give a point of reference, we also show the Most Frequent Class (MFC) baseline results.

Results. Table 6 shows that training on FEEL-IT brings stable performance even on datasets from different contexts. Note that the MFC accuracy is high because both datasets contain a wide range of emotions annotated as *anger*.

5 Related Work

Different works have explored emotion recognition approaches. However, few of them incorporate text in Italian. Indeed, currently, no general-purpose dataset for emotion recognition has been proposed for the Italian language. However, for Italian, there is a dataset for emotion recognition limited to Youtube and Facebook comments (Sprugnoli, 2020), and one for sentiment analysis SENTIPOLC16 (Barbieri et al., 2016). We used these datasets in the experimental evaluation to show that our model can also perform sentiment prediction.

Regarding other languages, Abdul-Mageed and Ungar (2017) proposes EmoNet, an English emotion dataset that has been collected using a keyword-based approach (e.g., tweets are retrieved using *#happy* as a marker for joy). The authors have obtained high accuracy with this dataset. Alternatevely, we approach the problem annotating manually and without using distant supervision. EmoTxt (Calefato et al., 2017) is an open-source toolkit for emotion prediction supporting prediction for different emotions for the English language: love, joy, surprise, anger, sadness, and fear. Nozza et al. (2017) propose a English corpus of tweets that comprises five different views for each message, i.e. subjective/objective, sentiment polarity, implicit/explicit, irony, emotion. Lamprinidis et al. (2021) introduce a novel dataset that covers multiple languages extracted from Facebook posts. Troiano et al. (2019) introduce a dataset in two languages, English and German, obtained through crowd-sourcing. Interestingly, Akhtar et al. (2019) propose a multi-model architecture that combines visual, auditory, and text information for both emotion and sentiment prediction in English.

6 Conclusions

We present FEEL-IT, a new corpus for emotion classification on Italian Twitter data, and release an open-source Python library to run both emotion and sentiment classification. Future work will focus on the extension of this dataset, considering other emotions and languages.

Acknowledgments

The authors are members of the Data and Marketing Insights research unit in the Bocconi Institute for Data Science and Analysis.

References

Muhammad Abdul-Mageed and Lyle Ungar. 2017. EmoNet: Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 718–728, Vancouver, Canada. Association for Computational Linguistics.

Zishan Ahmad, Raghav Jindal, Asif Ekbal, and Pushpak Bhattachharyya. 2020. Borrow from rich cousin: transfer learning for emotion detection using cross lingual embedding. *Expert Systems with Applications*, 139:112851.

Md Shad Akhtar, Dushyant Chauhan, Deepanway Ghosal, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Multi-task learning for multi-modal emotion recognition and sentiment analysis. 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)*, pages 370–379, Minneapolis, Minnesota. Association for Computational Linguistics.

Francesco Barbieri, Valerio Basile, Danilo Croce, M. Nissim, N. Novielli, and V. Patti. 2016. Overview of the Evalita 2016 SENTIment POLarity Classification Task. In *CLiC-it/EVALITA*.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Federico Bianchi, Silvia Terragni, Dirk Hovy, Debora Nozza, and Elisabetta Fersini. 2021. Cross-lingual Contextualized Topic Models with Zero-shot Learning. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*.

Lidong Bing, Sneha Chaudhari, Richard Wang, and William Cohen. 2015. Improving distant supervision for information extraction using label propagation through lists. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 524–529, Lisbon, Portugal. Association for Computational Linguistics.

Fabio Calefato, Filippo Lanubile, and Nicole Novielli. 2017. Emotxt: a toolkit for emotion recognition from text. In *Proceedings of the Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 79–80. IEEE.

Serina Chang, Ruiqi Zhong, Ethan Adams, Fei-Tzin Lee, Siddharth Varia, Desmond Patton, William Frey, Chris Kedzie, and Kathy McKeown. 2018. Detecting gang-involved escalation on social media using context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 46–56, Brussels, Belgium. Association for Computational Linguistics.

Shrey Desai, Cornelia Caragea, and Junyi Jessy Li. 2020. Detecting perceived emotions in hurricane disasters. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5290–5305, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Chunning Du, Haifeng Sun, Jingyu Wang, Qi Qi, and Jianxin Liao. 2020. Adversarial and domain-aware BERT for cross-domain sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4019–4028, Online. Association for Computational Linguistics.

Paul Ekman. 1992. An argument for basic emotions. *Cognition and Emotion*, 6(3-4):169–200.

Murhaf Fares, Andrey Kutuzov, Stephan Oepen, and Erik Velldal. 2017. Word vectors, reuse, and replicability: Towards a community repository of large-text resources. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 271–276, Gothenburg, Sweden. Association for Computational Linguistics.

Pere-Lluís Huguet Cabot, Verna Dankers, David Abadi, Agneta Fischer, and Ekaterina Shutova. 2020. The Pragmatics behind Politics: Modelling Metaphor, Framing and Emotion in Political Discourse. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4479–4488, Online. Association for Computational Linguistics.

Hamed Khanpour and Cornelia Caragea. 2018. Fine-grained emotion detection in health-related online posts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1160–1166, Brussels, Belgium. Association for Computational Linguistics.

Sotiris Lamprinidis, Federico Bianchi, Daniel Hardt, and Dirk Hovy. 2021. Universal Joy: A Data Set and Results for Classifying Emotions Across Languages. In *Proceedings of the Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Association for Computational Linguistics.

Elisa Leonardelli, Stefano Menini, and Sara Tonelli. 2020. DH-FBK @ haspeede2: Italian hate speech detection via self-training and oversampling. In *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), Online event, December 17th, 2020*, volume 2765 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Yosi Mass and Haggai Roitman. 2020. Ad-hoc document retrieval using weak-supervision with BERT and GPT2. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4191–4197, Online. Association for Computational Linguistics.

Saif Mohammad. 2012. #emotional tweets. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 246–255, Montréal, Canada. Association for Computational Linguistics.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, pages 1–17.

María Navas-Loro and Víctor Rodríguez-Doncel. 2019. Spanish corpora for sentiment analysis: a survey. *Language Resources and Evaluation*, pages 1–38.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2020. What the [MASK]? Making Sense of Language-specific BERT Models. *arXiv preprint arXiv:2003.02912*.

Debora Nozza, Elisabetta Fersini, and Enza Messina. 2017. A multi-view sentiment corpus. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 273–280, Valencia, Spain. Association for Computational Linguistics.

Debora Nozza, Pikakshi Manchanda, Elisabetta Fersini, Matteo Palmonari, and Enza Messina. 2021. LearningToAdapt with word embeddings: Domain adaptation of Named Entity Recognition systems. *Information Processing & Management*, 58(3):102537.

Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Languages Resources Association (ELRA).

Tommaso Pasini, Federico Scozzafava, and Bianca Scarlini. 2020. CluBERT: A cluster-based approach for learning sense distributions in multiple languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4008–4018, Online. Association for Computational Linguistics.

Nicole Peinelt, Dong Nguyen, and Maria Liakata. 2020. tBERT: Topic models and BERT joining forces for semantic similarity detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7047–7055, Online. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit.

2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2021. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020. With more contexts comes better performance: Contextualized sense embeddings for all-round word sense disambiguation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3528–3539, Online. Association for Computational Linguistics.

Rachele Sprugnoli. 2020. Multiemotions-it: a new dataset for opinion polarity and emotion analysis for italian. In *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.

Kevin Stowe, Michael J. Paul, Martha Palmer, Leysia Palen, and Kenneth Anderson. 2016. Identifying and categorizing disaster-related tweets. In *Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media*, pages 1–6, Austin, TX, USA. Association for Computational Linguistics.

Enrica Troiano, Sebastian Padó, and Roman Klinger. 2019. Crowdsourcing and validating event-focused emotion corpora for German and English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4005–4011.

Zhongqing Wang, Shoushan Li, Fan Wu, Qingying Sun, and Guodong Zhou. 2018. Overview of nlpcc 2018 shared task 1: Emotion detection in code-switching text. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 429–433. Springer.

A Data Statement

We follow Bender and Friedman (2018) on providing a Data Statement for the proposed FEEL-IT corpus.

Data has been annotated by two native Italian speakers, age group in 25-35, both with experience in computational linguistics. The data we share is not sensitive to personal information, as it does not contain information about individuals. Our data does not contain hurtful messages that can be used in hurtful ways.

B Additional results

As follows, we show the confusion matrices for UmBERTo-FT (Figure 1), UmBERTo-PT (Figure 2) and TF-IDF (Figure 3) representation models for the experiments in emotion recognition task with 10-fold cross validation on FEEL-IT.

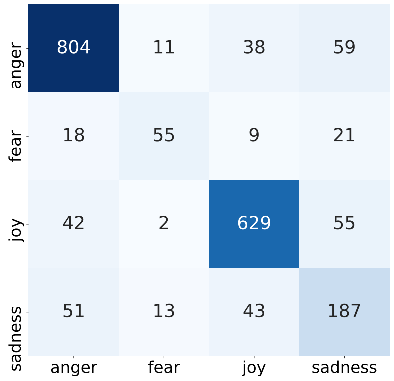

Figure 1: Confusion matrix of UmBERTo-FT predictions of cross-validated emotion classification on FEEL-IT.

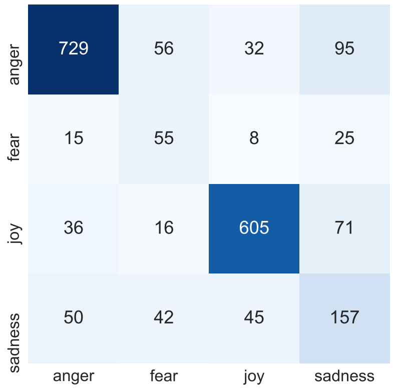

Figure 2: Confusion matrix of UmBERTo-PT predictions of cross-validated emotion classification on FEEL-IT.

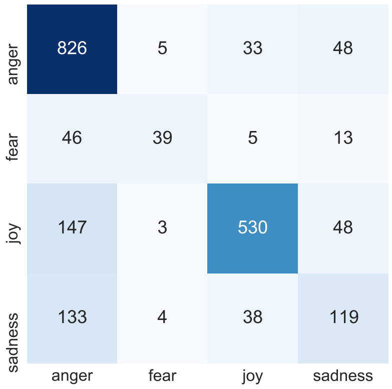

Figure 3: Confusion matrix of TF-IDF predictions of cross-validated emotion classification on FEEL-IT.