Can Emotion Carriers Explain Automatic Sentiment Prediction? A Study on Personal Narratives

Seyed Mahed Mousavi, Gabriel Roccabruna, Aniruddha Tammewar, Steve Azzolin, Giuseppe Riccardi
Signals and Interactive Systems Lab, University of Trento, Italy
{mahed.mousavi,gabriel.roccabruna,giuseppe.riccardi}@unitn.it

Abstract

Deep Neural Networks (DNN) models have achieved acceptable performance in sentiment prediction of written text. However, the output of these machine learning (ML) models cannot be natively interpreted. In this paper, we study how the sentiment polarity predictions by DNNs can be explained and compare them to humans' explanations. We crowdsource a corpus of Personal Narratives and ask human judges to annotate them with polarity and select the corresponding token chunks - the Emotion Carriers (EC) - that convey narrators' emotions in the text. The interpretations of ML neural models are carried out through Integrated Gradients method and we compare them with human annotators' interpretations. The results of our comparative analysis indicate that while the ML model mostly focuses on the explicit appearance of emotions-laden words (e.g. happy, frustrated), the human annotator predominantly focuses the attention on the manifestation of emotions through ECs that denote events, persons, and objects which activate narrator's emotional state.

1 Introduction

Neural data-driven models have managed to perform comparably well in various tasks related to natural language processing (Eberts and Ulges, 2020; Adoma et al., 2020). Nevertheless, the definition and the training processes of such models have made their decision non-natively interpretable. Several studies and experiments have been conducted to address this issue and explain the decision outputs of such models in various tasks such as emotion prediction (Yang et al., 2019), question answering (Ramnath et al., 2020), the classification of linguistic styles (Hayati et al., 2021), and lexicon-based sentiment prediction (Hwang and Lee, 2021).

Sentiment analysis is a well-established field of research that aims to extract sentiment and its aspects in a written text. Its performances have reached acceptable levels in different domains such as product reviews (Xie et al., 2020), movie reviews (Thongtan and Phienthrakul, 2019), social media (Tam et al., 2021), financial news (Takala et al., 2014), and Personal Narratives (PN) which are recollections of real-life events that are experienced by the narrator (Tammewar et al., 2019).

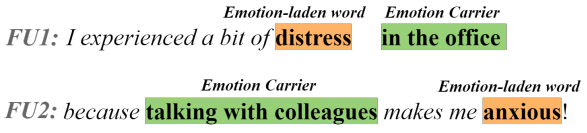

Figure 1: Example of a sentence consisting of two Functional Units (FU1, FU2), the basic units of annotation. Emotion-laden words in each Functional Unit manifest a sentiment explicitly while Emotion Carriers describe the events, persons or objects conveying emotions.

Recently, a deeper understanding of the expressed sentiment and emotion has gained growing research interest (Tammewar et al., 2020, 2021; Bayerl et al., 2021; Ding et al., 2020). These works focus on a more fine-grained analysis on the expressed sentiment/emotion by identifying the Emotion Carriers (entities or actions that explain, cause or carry the emotion). The concept of Emotion Carriers (EC) was first introduced by Tammewar et al. (2020) for German PNs. In this genre of text, the identification of ECs may help in better understanding the emotional state of the narrator and what has caused distress (Tammewar et al., 2021; Bayerl et al., 2021).

In this work, we address the problem of analyzing and comparing the text chunks used by machines and humans when predicting the sentiment polarity of text documents. For this study we have selected the Personal Narrative genre since it is rich with entities and relations which are sparsely distributed. We identify the tokens that contribute to the model's prediction according to their attributions given by Integrated Gradients (Sundarara-

jan et al., 2017), an Explainable-AI technique, and compare them with the tokens tagged as ECs by the human annotator. Our comparative analysis shows the human annotator identifies the tokens that explain an event or its participants as the carrier of emotions and sentiments, which clearly convey the activation of the emotional state in the narrator, even though they are not explicitly manifesting a sentiment. Meanwhile, the DNN model bases its decision mostly on a limited set of tokens which belong to the category of emotion-laden words (see Figure 1 for an example).

We summarize our contribution as follows:

- The annotation of a dataset of Personal Narratives to obtain the sentiment polarity, and the Emotion Carriers at the Functional Unit (Bunt et al., 2010) level to take into account the communicative functions. This is in contrast with traditional annotation at the document or sentence level.
- The evaluation of the annotation results and training a sentiment prediction model based on the AlBERTo architecture (Polignano et al., 2019) using the annotated data, as well as a baseline architecture for the task of Emotion Carrier Detection.
- The study of the tokens contributing to the model's prediction of sentiment and comparing them with the Emotion Carriers identified by the human annotator, and the contribution of the Emotion Carriers in the prediction of the model by their influence on the output confidence score.

2 Literature Review

AI Explainability There have been several interesting works to address the unexplainability of neural architectures. Danilevsky et al. (2020) conducted a survey study on explainable AI (XAI) in natural language processing, summarizing the various XAI methods used by researchers. Bodria et al. (2020) proposed an attention model to investigate the words that contribute to the sentiment prediction, by adding an additional attention layer on top of the BERT architecture to fuse the token embeddings in one vector used to compute the prediction. Bacco et al. (2021) used the attention weights technique to extract summaries of reviews to explain the sentiment prediction of a Transformer-based model, by using a simplified model with 2 layers and one attention head per layer. Torres et al. (2021) designed a deep neural network with an interpretable decision process to recognize emotions from the Electroencephalography (EEG) signals.

While the approaches based on attention weights require a change in the architecture of the model, LIME (Local Model-Agnostic Explanations) (Ribeiro et al., 2016) and the Integrated Gradients technique (Sundararajan et al., 2017) can be applied to any model without changing the architecture. Using LIME, Hwang and Lee (2021) extracted a sentiment lexicon used as a weak classifier to categorize unseen examples to augment the initial training set. Similarly, Carton et al. (2018) used LIME and hard-attention to extract spans of text that convey personal attacks. Furthermore, Hayati et al. (2021) used the Integrated Gradients to compare most relevant tokens for the human and the machine in predicting the linguistic style of a text.

Emotion & Sentiment Analysis An approach to perform fine-grained analysis on the expressed emotion in the text is the task of emotion cause extraction (Chen et al., 2018; Xia and Ding, 2019; Ding et al., 2020; Gui et al., 2016). The aim of this task is to identify the explicit or implicit expressions of emotions in the text, as well as the corresponding causes or triggers of the emotion as a span in the text (Turcan et al., 2021; Li et al., 2021a,b). However, most of the works on this task have focused on datasets of news (Bostan et al., 2020; Gui et al., 2016) and microblogs (Oberländer and Klinger, 2020), which are very different from Personal Narratives.

Understanding of Personal Narratives (PN) is a comparatively new domain and is gaining growing attention in the research community (Stappen et al., 2019; Tammewar et al., 2019; Schuller et al., 2018; Rathner et al., 2018; Ong et al., 2021). Compared to the mentioned genres of text, PNs have a different and more complex structure as they are personal recollection of real-life events and may involve multiple characters, and several sub-events (Mousavi et al., 2021; Tammewar et al., 2019). A stream of works has been carried out on the fine-grained emotion analysis of PNs that tries to capture the semantics of the emotions through Emotion Carriers (EC), including the annotation of ECs (Tammewar et al., 2020) as well as the automatic recognition of the ECs (Tammewar et al., 2021; Bayerl et al.,

2021). In these works, every PN is associated with a positive or negative emotion and the ECs are defined as the persons, objects or actions that explain the emotion felt by the narrator, after recollecting the event.

3 Data Collection & Annotation

We used an extended version of the dataset of PNs from users receiving Cognitive Behavioural Therapy to handle their distress more effectively, introduced previously by Mousavi et al. (2021). Each PN encompasses a real-life personal event that has activated the narrator's emotional state, the participants of the event as well as the details about the user's thought and emotions. During two periods of 3 months, we collected 481 personal narratives written by 45 Italian speaker users, with the average length of 51 tokens per narrative and overall dictionary size of 5875 tokens.

3.1 Annotation of Sentiment & Emotion Carriers

We annotate the obtained dataset of PNs, with the sentiment and the Emotion Carrier tokens for each narrative[1]. The mentioned studies on identifying ECs (Tammewar et al., 2020; Bayerl et al., 2021) focus on the identification of emotion and the corresponding ECs at the narrative level. However, in this work we conduct a deeper analysis and identify the emotion and the corresponding ECs for each Functional Unit of the PN, making it possible to capture the emotion changes of the narrator throughout the narrative. A Functional Unit (FU) is defined as a minimal contiguous span in the text that represents coherent communicative intention (Bunt et al., 2010). We segment each PN to its FUs, using a RoBERTa-based model[2] (Liu et al., 2019), fine-tuned on ISO standard Dialogue Act tagging in Italian (Roccabruna et al., 2020) to jointly perform FU segmentation and Dialogue Act tagging. As the result, we obtained 4273 FUs to be annotated (approximately equal to 9 FUs for each narrative on average).

We recruited 3 Italian native speaker annotators from a pool of graduate students based on their research interests and previous experience with data annotation. The annotators were asked to annotate the sentiment polarity of the FU using a 5-point bipolar scale from -2 ($unpleasant$) to 2 ($pleasant$) with 0 representing $neutral$. The annotators were asked to adopt the point of view of the narrator. In the cases where the sentiment of the FU was not clear by its content, the annotators were asked to consider the adjacent FUs as context for better understanding.

For the FUs with an assigned sentiment polarity of positive or negative, the annotators were further asked to select the ECs that convey and carry the annotated sentiment of the narrator in the corresponding FU. Considering the characteristics of PNs as the recollection of real-life events, we focused on the manifestations of the sentiment in terms of persons, objects, places, organizations or actions that affected the narrator's emotional state. Therefore, we provided the annotators with a list of noun-chunks and verb-chunks in the FU as EC-candidate spans to select from, and excluded the explicit emotion-laden words such as *happy, sad, enjoyed, and overwhelmed*, since they directly express certain sentiment polarity. Besides, this approach helped to reduce the cognitive load of the subtask.

Prior to the annotation, we carried out a training session for the annotators administered by a psychotherapist, followed by two training batches by which a satisfactory Inter-Annotator Agreement (IAA) was achieved (the results of the training batches were manually controlled and few adjustments were made with the annotators and to the guidelines). We then distributed the samples in 10 batches with 20% overlap in each batch annotated by all 3 annotators (to monitor the IAA and ensure the annotation quality) and the remaining 80% annotated by a single annotator.

3.2 Annotation Results Analysis

Using the 481 Personal Narratives, we annotated 4273 functional units[3]. As the results, the majority of the FUs, 60%, were annotated as neutral, while 13% and 27% of them were labeled as positive and negative respectively. The Inter-Annotator Agreement (IAA), computed with the Fleiss' κ coefficient (Fleiss, 1971), on the sentiment annotation is 0.67 (Substantial) on the 5-point scale results, and 0.73 (Substantial) on the 3-point scale (obtained by regrouping the values into three groups of *positive*

[1] We are currently applying for further funds to anonymize the corpus and publish a version of the corpus that respects users' privacy and deontological requirements.

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

[3] As example of valence and ECs annotation on a PN at the level of Functional Units: https://gitlab.com/sislab/PNs_Val-EC_annotation

Polarity	Freq.	EC	non-EC
Positive	13%	566 (28%)	736 (30%)
Negative	27%	1425 (72%)	1725 (70%)
Neutral	60%	-	-

Table 1: The distribution of polarity and Emotion Carriers (EC) in the annotated dataset of Personal Narrative at functional unit level.

Model	F1	Prec.	Rec.
bi-LSTM + attn.	0.66	0.70	0.66
AlBERTo Emb.	0.69	0.69	0.69
AlBERTo Emb.+$[CLS]$	**0.70**	**0.70**	**0.70**

Table 2: Results of EC Detection experiments on the test set. All scores are measured with the "macro" average strategy. The AlBERTo-based architecture with the concatenation of $[CLS]$ token achieves the best performance.

{1,2}, *negative {-2,-1}* and *neutral {0}*). Furthermore, the IAA on the examples that were labelled with a non-neutral polarity by all annotators is 0.98 (Almost Perfect).

Regarding the EC selection, out of 4452 EC-candidate spans in the FUs that were labeled with a non-neutral sentiment polarity, 1991 spans (45%) were chosen as EC by the annotators, resulting in 2551 EC tokens (tokens in the EC-span) and the EC dictionary size of 962. The IAA on the EC annotation is 0.4 (Fair), computed by considering each EC-candidate as an example to annotate where the labels are *yes* if it is an EC, and *no* otherwise.

The statistics regarding the labelled ECs and the sentiment distribution are presented in Table 1. For our experiments, we split the obtained annotated dataset into training (80%), validation (10%) and test (10%) sets, stratified on the polarity distribution and on the lengths of the PN.

3.3 Emotion Carrier Detection Baseline

We trained a baseline model to assess the EC annotation on the PN dataset for the task of EC detection. The approaches used in previous works (Tammewar et al., 2021; Bayerl et al., 2021) do not fit with our case, since the annotators were asked to select the EC from a predefined set of candidates, rather than selecting any token in the text. Thus, in our case the model is tasked to classify each EC-candidate span as EC or non-EC.

The first part of the architecture computes the tokens embedding of each FU. Afterwards, we extract the encoded representation of the EC-candidate tokens and perform max-pooling, which takes the maximum value for every dimension of the vector encoding, producing the vector representation of the EC-candidate. The vector representation is then given as input to the classification layer (dense layer + softmax) yielding the probability distribution over the EC and non-EC classes. To compute the embeddings, we experimented with bi-LSTM with attention and AlBERTo, a pre-trained BERT-based model for the Italian language (Polignano et al., 2019). In the experiments with the AlBERTo model, we experimented concatenating the representation of the $[CLS]$ token with the EC-candidate representation, to better consider the context during the classification.

The results of these experiments, summarized in Table 2, indicate that the outperforming baseline combination is obtained by using the AlBERTo model for the input representation with the concatenation of the $[CLS]$ token.

3.4 Sentiment Prediction Model

We trained a sentiment prediction model to predict the polarity at the level of functional units. Our model is based on the AlBERTo architecture (Polignano et al., 2019) with a three-heads output layer, instead of the original two-heads fully connected layers, to predict the sentiment polarity of each FU over the 3-label output space of *negative, positive* and *neutral*. We split the training set of the SENTIPOLC16 dataset (Barbieri et al., 2016)[4] into training and validation sets of 90% and 10%, in a stratified manner. We then used the training set to fine-tune the model in the first step, and the validation set in the next step for hyper-parameter optimization and selecting the best model using the Optuna framework (Akiba et al., 2019). Using the obtained hyper-parameters[5], the model was then further fine-tuned on our own collected dataset of annotated functional units extracted from PNs. The results of these experiments are presented in Table 3.

[4]SENTIPOLC16 is a dataset of tweets in the Italian language

[5]`learning_rate=6.599e-05, weight_decay=0.0215, warmup_steps=0.899, num_epochs=11`

Model	F1	Prec.	Rec.
AlBERTo_SP16	0.64	0.63	0.70
AlBERTo_opt_SP16	0.63	0.62	0.71
AlBERTo_opt_SP16+PN	**0.76**	**0.76**	**0.76**

Table 3: Macro F1, Precision, and Recall of the sentiment prediction models optimized in different settings. *AlBERTo_SP16* is the vanilla AlBERTo model fine-tuned on SENTIPOLC16; *AlBERTo_opt_SP16* is the model optimized utilizing validation split; and *AlBERTo_opt_SP16+PN* is the *AlBERTo_opt_SP16* further fine-tuned on the training set of our Personal Narratives dataset. All evaluation results are obtained using the test split of the Personal Narratives dataset.

4 Prediction Decision Explainability

We investigate the explainability of the automatic sentiment prediction by comparing the tokens influencing the prediction with those selected by the human judge as ECs. In order to detect the tokens crucial to the model's prediction, we use the attribution assigned to each token by the Integrated Gradients (Sundararajan et al., 2017) technique. Integrated Gradients (IntGrad) is an attribution method for Explainable AI which builds on top of the classic backward gradient analysis. Given our sentiment prediction model $f(FU)$, where FU is the functional unit $FU = \{w_1, w_2, .., w_n\}$ and $w_i \in R^d$ are the token embeddings , the backward gradient is given by:

$$BackwardGrad_j(w_i) = \frac{\partial f}{\partial w_{ij}} \quad (1)$$

measuring how much perturbing the input token w_i by an infinitesimal amount along dimension j affects the output of function f. The IntGrad method extends this by computing the integral of the derivative along the path connecting a baseline token w', which is a neutral element, to the input point w:

$$IntGrad_j(w_i) = (w_{ij} - w') \int_{\alpha=0}^{1} \frac{\partial f(w' + \alpha(w_{ij} - w'))}{\partial w_{ij}} d\alpha \quad (2)$$

where $\alpha \in [0, 1]$ draws a linear path, from the baseline token to the input token, along which the gradients are integrated. In our studies, we used a zero vector for the baseline token w', and the open-source library Captum (Kokhlikyan et al., 2020) for efficient IntGrad computation. In cases that a token is split into several subtokens by the tokenizer of our model (Kudo and Richardson, 2018), we average the Integrated Gradients attributions of the subtokens, to get the attribution of the whole token.

4.1 Token Analysis based on IntGrad Attributions

Using the test set samples for which the model predicts the sentiment polarity correctly, we employ two approaches regarding the explainability analysis. In the first approach, we extract the tokens influential or crucial to the prediction process of the model based on their Integrated Gradients (IntGrad) attributions, and study whether or not they belong to the spans annotated as EC by the human annotator.

In order to identify tokens crucial to the model's prediction we experimented with two different thresholds for the IntGrad attribution:

- **Greater than 0 (G0)**: This baseline is based on the fact that each token with a positive IntGrad attribution value has a positive influence on the prediction. Nevertheless, tokens with small IntGrad attributions have a marginal contribution and thus they are noisy for our analysis;
- **Lower Bound (LB)**: This threshold is obtained uniquely for each FU and is measured by consecutively masking each token in the FU, with a zero-vector embedding, in a descending order of IntGrad attributions until a change in the polarity prediction is observed. The IntGrad attribution of the last masked out token is then selected as the LB threshold.

The results of this analysis using the two mentioned threshold policies are presented in Table 4 and Figure 2. The analysis indicates that although 67.9% of the EC tokens (tokens in ECs selected by human annotators) have a positive contribution to the model's prediction, more than 60% of the tokens with an attribution above the thresholds do not overlap with the EC tokens. Nevertheless, the majority of EC tokens with an attribution higher than the thresholds are EC-heads, regardless of the threshold policy. Furthermore, the distributions of the Content Words (CW), i.e. nouns, verbs and adjectives, confirm our previous assumption that **G0** threshold is noisy since 54% of tokens above this threshold are non-CWs, while this number is smaller than 20% for the tokens with an IntGrad attribution higher than the **LB**. The CWs in LB and G0 groups are distributed as 52% nouns, 27%

Threshold (Thr.)	G0	LB
Tokens with IntGrad A.>Thr.	482 (46% CW)	109 (81% CW)
Tokens w. IntGrad A.>Thr. in EC-span	141 29.3%	43 39.5%
Tokens w. IntGrad A.>Thr. that are EC-heads	82 18.1%	32 29.3%

Table 4: The analysis of tokens influencing the model's prediction based on two different policies for the IntGrad attribution (IntGrad A.), namely Greater than 0 (G0) and Lower Bound (LB). Regardless of the threshold policy, the tokens inside the EC-span that contribute to the model's prediction are less than 40%.

Token set	Positive	Negative	Neutral
CW in G0	3.9%	13.6%	82.5%
CW in LB	10.3%	29.9%	59.8%
CW EC tokens	0.7%	4.0%	95.3%

Table 5: The polarity distribution of the Content Words (CW) with IntGrad attribution higher than the different thresholds. The results indicate that the majority of CWs in EC tokens are neutral and they do not represent any emotions explicitly. The polarity was retrieved using the OpeNER sentiment lexicon for the Italian language.

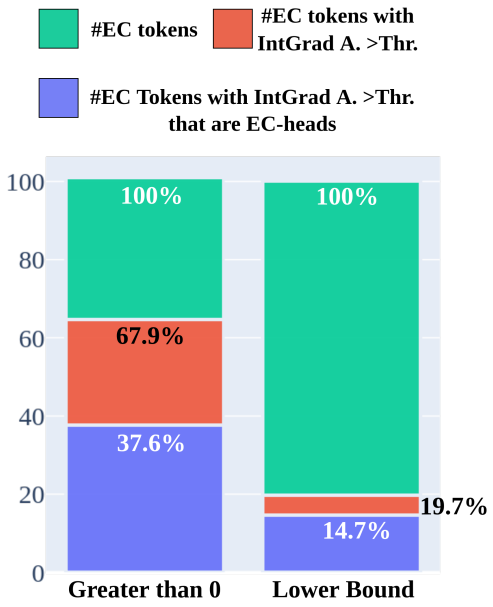

Figure 2: The percentage of the tokens in EC-spans with an Integrated Gradient attribution (IntGrad A.) higher than the threshold (Thr.). The majority of EC tokens with an attribution higher than the Lower Bound are EC-heads.

verbs, 21% adjectives, and 47% nouns, 40% verbs and 13% adjectives, respectively.

In the next step, we further analyzed the polarity distribution of CWs by using the OpeNER[6] lexicon-based sentiment model. The results, presented in Table 5, show that the percentage of non-neutral CWs in the ECs is less than 5%, while more than 40% of the influential tokens, i.e. tokens with attributions over the **LB** threshold, represent a positive or negative polarity. This remarks the importance of emotion-laden words, such as *anxiety*, *fear* and *worry*, for the model in predicting the sentiment, and suggests that the model mostly focuses on the tokens that explicitly convey emotions, and the ECs (as the implicit manifestations of emotions) are less significant in its decision process.

4.2 Contribution of ECs to the Model's Decision

For the second approach, we evaluate the influence of the ECs selected by the human annotators in the decision process of the model. For this purpose, we mask out the EC-span in the Functional Unit with the highest IntGrad attribution, and measure the drop in the confidence score for the initially predicted polarity. The confidence score represents the probability assigned by the model to a given class, which in our case the classes can be either *positive* or *negative*. In the next step, we extend this analysis to the token level and measure the drop in the confidence score caused by masking out the EC-head with the highest IntGrad attribution, as well as all EC-heads present in the corresponding FU.

The results, shown in Table 6, present the strong contribution of emotion-laden words that explicitly manifest the sentiment on the model's decision. Furthermore, the confidence drop caused by masking the EC-span is higher than masking only the head of the corresponding EC, suggesting that all the tokens in the EC-span contribute to the prediction confidence. However, the highest drop is achieved by masking the most influential token (the token with the highest IntGrad attribution) and emotion-laden words, respectively. These results once again support the findings of the previous analysis, suggesting the importance of tokens that explicitly manifest a sentiment in the decision process of the model.

[6] https://www.opener-project.eu/, This publicly available lexicon was semi-automatically created starting from 1,000 manually controlled keywords

Masked Content in FU	Conf. Score Drop
EC-Span w. highest IntGrad A.	0.15
EC-Head w. highest IntGrad A.	0.09
EC-Heads in FU	0.14
Token w. highest IntGrad A.	0.55
Emotion-laden Words	0.36

Table 6: The drop in the confidence score of the predicted polarity caused by masking out selected contents in Functional Units. The results show that the Emotion-laden words have a stronger influence than the tokens selected as ECs by the human annotator.

5 Conclusion

In this work we studied whether the sentiment prediction decision of DNN models can be explained by Emotion Carriers, spans of text that convey and carry emotions. We have focused our study on Personal Narratives which encompass real-life events and experiences that activate the emotional state of the narrator. We have collected a dataset of Personal Narratives and conducted an annotation task for sentiment polarity and Emotion Carrier selection at the Functional Unit for each narrative. We have then developed a sentiment prediction model based on AlBERTo architecture (Polignano et al., 2019). We have investigated whether the decision of the model is based on the Emotion Carriers that the human annotator selected to explain the sentiment of the text. Furthermore, we have studied the impact of the Emotion Carriers on the confidence score of the polarity prediction model. Our analysis has shown that the human annotators tend to focus on manifestation of emotions through words describing actions and events that have activated the emotional state of the narrator. However, the model bases its decision on explicit representations of sentiment such as emotion-laden words.

Acknowledgements

The research leading to these results has received funding from the European Union – H2020 Programme under grant agreement 826266: COADAPT.

References

Acheampong Francisca Adoma, Nunoo-Mensah Henry, and Wenyu Chen. 2020. Comparative analyses of bert, roberta, distilbert, and xlnet for text-based emotion recognition. In *2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, pages 117–121. IEEE.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Luca Bacco, Andrea Cimino, Felice Dell'Orletta, and Mario Merone. 2021. Extractive summarization for explainable sentiment analysis using transformers. In *DeepOntoNLP/X-SENTIMENT@ESWC*.

Francesco Barbieri, Valerio Basile, Danilo Croce, Malvina Nissim, Nicole Novielli, and Viviana Patti. 2016. Overview of the evalita 2016 sentiment polarity classification task. In *Proceedings of third Italian conference on computational linguistics (CLiC-it 2016) & fifth evaluation campaign of natural language processing and speech tools for Italian. Final Workshop (EVALITA 2016)*.

Sebastian P. Bayerl, Aniruddha Tammewar, Korbinian Riedhammer, and Giuseppe Riccardi. 2021. Detecting emotion carriers by combining acoustic and lexical representations. In *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2021, Cartagena, Colombia, December 13-17, 2021*, pages 31–38. IEEE.

Francesco Bodria, André Panisson, Alan Perotti, and Simone Piaggesi. 2020. Explainability methods for natural language processing: Applications to sentiment analysis. In *SEBD*.

Laura Ana Maria Bostan, Evgeny Kim, and Roman Klinger. 2020. GoodNewsEveryone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1554–1566, Marseille, France. European Language Resources Association.

Harry Bunt, Jan Alexandersson, Jean Carletta, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Kiyong Lee, Volha Petukhova, Andrei Popescu-Belis, Laurent Romary, et al. 2010. Towards an iso standard for dialogue act annotation. In *Seventh conference on International Language Resources and Evaluation (LREC'10)*.

Samuel Carton, Qiaozhu Mei, and Paul Resnick. 2018. Extractive adversarial networks: High-recall explanations for identifying personal attacks in social media posts. In *EMNLP*.

Ying Chen, Wenjun Hou, Xiyao Cheng, and Shoushan Li. 2018. Joint learning for emotion classification and emotion cause detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 646–651.

Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A survey of the state of explainable ai for natural language processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 447–459.

Zixiang Ding, Rui Xia, and Jianfei Yu. 2020. Ecpe-2d: emotion-cause pair extraction based on joint two-dimensional representation, interaction and prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3161–3170.

Markus Eberts and Adrian Ulges. 2020. Span-based joint entity and relation extraction with transformer pre-training. In *ECAI 2020*, pages 2006–2013. IOS Press.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Lin Gui, Dongyin Wu, Ruifeng Xu, Qin Lu, and Yu Zhou. 2016. Event-driven emotion cause extraction with corpus construction. In *EMNLP*.

Shirley Hayati, Dongyeop Kang, and Lyle Ungar. 2021. Does bert learn as humans perceive? understanding linguistic styles through lexica. pages 6323–6331.

Hohyun Hwang and Younghoon Lee. 2021. Semi-supervised learning based on auto-generated lexicon using xai in sentiment analysis. In *RANLP*.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. 2020. Captum: A unified and generic model interpretability library for pytorch.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71.

Xiangju Li, Wei Gao, Shi Feng, Daling Wang, and Shafiq Joty. 2021a. Span-level emotion cause analysis by bert-based graph attention network. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 3221–3226.

Xiangju Li, Wei Gao, Shi Feng, Yifei Zhang, and Daling Wang. 2021b. Boundary detection with bert for span-level emotion cause analysis. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 676–682.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Seyed Mahed Mousavi, Alessandra Cervone, Morena Danieli, and Giuseppe Riccardi. 2021. Would you like to tell me more? generating a corpus of psychotherapy dialogues. In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pages 1–9, Online. Association for Computational Linguistics.

Laura Ana Maria Oberländer and Roman Klinger. 2020. Token sequence labeling vs. clause classification for english emotion stimulus detection. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 58–70.

Desmond C. Ong, Zhengxuan Wu, Zhi-Xuan Tan, Marianne Reddan, Isabella Kahhale, Alison Mattek, and Jamil Zaki. 2021. Modeling emotion in complex stories: The stanford emotional narratives dataset. *IEEE Transactions on Affective Computing*, 12(3):579–594.

Marco Polignano, Pierpaolo Basile, Marco de Gemmis, Giovanni Semeraro, and Valerio Basile. 2019. AlBERTo: Italian BERT Language Understanding Model for NLP Challenging Tasks Based on Tweets. In *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019)*, volume 2481. CEUR.

Sahana Ramnath, Preksha Nema, Deep Sahni, and Mitesh M. Khapra. 2020. Towards interpreting BERT for reading comprehension based QA. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3236–3242, Online. Association for Computational Linguistics.

Eva-Maria Rathner, Yannik Terhorst, Nicholas Cummins, Björn Schuller, and Harald Baumeister. 2018. State of mind: Classification through self-reported affect and word use in speech. *Proc. Interspeech 2018*, pages 267–271.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*.

Gabriel Roccabruna, Alessandra Cervone, and Giuseppe Riccardi. 2020. Multifunctional iso standard dialogue act tagging in italian. *Seventh Italian Conference on Computational Linguistics (CLiC-it)*.

Björn Schuller, Stefan Steidl, Anton Batliner, Peter B. Marschik, Harald Baumeister, Fengquan Dong, Simone Hantke, Florian B. Pokorny, Eva-Maria Rathner, Katrin D. Bartl-Pokorny, Christa Einspieler, Dajie Zhang, Alice Baird, Shahin Amiriparian, Kun Qian, Zhao Ren, Maximilian Schmitt, Panagiotis Tzirakis, and Stefanos Zafeiriou. 2018. The interspeech

2018 computational paralinguistics challenge: Atypical self-assessed affect, crying heart beats. In *Proc. Interspeech 2018*, pages 122–126.

Lukas Stappen, Nicholas Cummins, Eva-Maria Meßner, Harald Baumeister, Judith Dineley, and Björn Schuller. 2019. Context modelling using hierarchical attention networks for sentiment and self-assessed emotion detection in spoken narratives. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6680–6684.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks.

Pyry Takala, Pekka Malo, Ankur Sinha, and Oskar Ahlgren. 2014. Gold-standard for topic-specific sentiment analysis of economic texts. In *LREC*.

Sakirin Tam, Rachid Ben Said, and Ö Özgür Tanriöver. 2021. A convbilstm deep learning model-based approach for twitter sentiment classification. *IEEE Access*, 9:41283–41293.

Aniruddha Tammewar, Alessandra Cervone, Eva-Maria Messner, and Giuseppe Riccardi. 2019. Modeling User Context for Valence Prediction from Narratives. In *Proc. Interspeech 2019*, pages 3252–3256.

Aniruddha Tammewar, Alessandra Cervone, Eva-Maria Messner, and Giuseppe Riccardi. 2020. Annotation of emotion carriers in personal narratives. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1517–1525.

Aniruddha Tammewar, Alessandra Cervone, and Giuseppe Riccardi. 2021. Emotion Carrier Recognition from Personal Narratives. In *Proc. Interspeech 2021*, pages 2501–2505.

Tan Thongtan and Tanasanee Phienthrakul. 2019. Sentiment classification using document embeddings trained with cosine similarity. In *ACL*.

Juan Manuel Mayor Torres, Mirco Ravanelli, Sara E. Medina-DeVilliers, Matthew Daniel Lerner, and Giuseppe Riccardi. 2021. Interpretable sincnet-based deep learning for emotion recognition from eeg brain activity. *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 412–415.

Elsbeth Turcan, Shuai Wang, Rishita Anubhai, Kasturi Bhattacharjee, Yaser Al-Onaizan, and Smaranda Muresan. 2021. Multi-task learning and adapted knowledge models for emotion-cause extraction. *arXiv e-prints*, pages arXiv–2106.

Rui Xia and Zixiang Ding. 2019. Emotion-cause pair extraction: A new task to emotion analysis in texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1003–1012.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 33:6256–6268.

Yang Yang, Deyu Zhou, Yulan He, and Meng Zhang. 2019. Interpretable relevant emotion ranking with event-driven attention. In *EMNLP*.