CARES: CAuse Recognition for Emotion in Suicide notes

Soumitra Ghosh[1][0000−0003−1910−4320], Swarup Roy[2], Asif Ekbal[1][0000−0003−3612−8834], and Pushpak Bhattacharyya[1][0000−0001−5319−5508]

[1] Indian Institute of Technology Patna, India
[2] National Institute of Technology, Durgapur, India
{ghosh.soumitra2,shahebroy7}@gmail.com,asif@iitp.ac.in,pb@cse.iitb.ac.in

Abstract. Inspired by recent advances in emotion-cause extraction in texts and its potential in research on computational studies in suicide motives and tendencies and mental health, we address the problem of *cause identification* and *cause extraction* for emotion in suicide notes. We introduce an emotion-cause annotated suicide corpus of 5769 sentences by labeling the benchmark CEASE-v2.0 dataset (4932 sentences) with causal spans for existing annotated emotions. Furthermore, we expand the utility of the existing dataset by adding emotion and emotion cause annotations for an additional 837 sentences collected from 67 non-English suicide notes (Hindi, Bangla, Telugu). Our proposed approaches to emotion-cause identification and extraction are based on pre-trained transformer-based models that attain performance figures of 83.20% accuracy and 0.76 Ratcliff-Obershelp similarity, respectively. The findings suggest that existing computational methods can be adapted to address these challenging tasks, opening up new research areas.

Keywords: emotion cause, suicide notes, XLM-R, BERT, SpanBERT

1 Introduction

Suicide continues to be one of the major causes of death across the world. Suicide rates have risen by 60% globally in the previous 45 years[3]. Suicide is currently one of the top three causes of mortality for those aged 15 to 44. Both men and women commit suicide at higher rates throughout Europe, notably in Eastern Europe. India and China account for about 30% of all suicides globally[4]. By writing a suicide note, a person communicates sentiments that would otherwise lie covert and fester. Suicide notes may be used as both an explanation and a therapeutic tool to help family members comprehend the suicide [10].

In this work, we primarily aim at achieving two objectives:

– *Produce a gold standard corpus annotated with causal spans for emotion annotated sentences in suicide notes.*
– *Develop a benchmark setup for emotion cause recognition in suicide notes, specifically, cause identification and cause extraction.*

[3] https://www.who.int/news/item/17-06-2021-one-in-100-deaths-is-by-suicide
[4] https://www.befrienders.org/suicide-statistics

Given the limits of automated approaches in suicidal research, this study will aid the research community by introducing a publicly available emotion cause annotated corpus of 5769 sentences from suicide notes (whose availability is otherwise scarce). Additionally, the attempt to make the available benchmark CEASE-v2.0 dataset [8] multilingual, by adding 17% new sentences from 67 non-English suicide notes (Hindi, Bangla and Telugu) will increase the utility of the CEASE-v2.0 dataset. Lastly, the proposed evaluation methods on the dataset can serve as solid baselines for future research on this dataset and related topics.

2 Background

The Emotion Cause Extraction (ECE) task aims to identify the possible causes of certain emotions from text. The basic premise is that such phrases are good descriptors of the underlying causes of the expressed emotions [19]. When compared to emotion classification, this is a far more challenging problem. In the early 1960s, several scholars [1,20] began analyzing the content of suicide notes to investigate the various reasons for suicide. The socioeconomic and psychological causes of suicides were examined in [17].

Emotion cause analysis has received greater attention in recent years in the sentiment analysis and text mining fields [16,6]. The work in [9] introduced an emotion cause annotated dataset built from SINA[5] news for event-driven emotion cause extraction. Deep learning has lately piqued the interest of the ECE community. The authors in [2] proposed a joint neural network-based method for emotion extraction and emotion cause extraction, capturing mutual benefits across these two emotion analysis tasks. In [21], a two-step technique was designed to solve the related job of emotion cause pair extraction (ECPE). Emotion identification and cause extraction was conducted first by a multitask framework, followed by emotion-cause pairing and filtering. To tackle the emotion cause pair extraction challenge, [3] suggested a graph neural network-based solution. Recently, [14] introduced the task of recognizing emotion cause in conversations (RECCON) and presented the emotion cause annotated RECCON dataset.

Despite numerous studies on ECE in other fields, no such study on suicide research utilizing computational techniques exists to our knowledge. Our effort aims at bridging this gap by performing ECE in suicide notes. We consider the CEASE-v2.0 [8] suicide notes corpus, which is an improved version of the CEASE dataset provided in the introductory paper [7] by the same authors.

3 Corpus Development

Following the methods for data collection and annotation in [7,8], we introduce multilingual CARES_CEASE-v2.0 corpus with emotion cause annotations.

Data Collection: We collected 67 suicide notes, from various Indian news websites such as 'patrika.com', 'jkstudenttimes.com', etc., of three most popularly spoken languages in India, Hindi (46 notes), Bengali (15 notes) and Telugu (6 notes). After proper anonymization, each note was sentence tokenized and the resultant bag of sentences were shuffled for reconstruction of the actual notes.

[5] http://news.sina.com.cn/society/

Annotations: Three annotators (two undergraduate students and one doctoral researcher of computer science discipline) with sufficient subject knowledge and experience on construction of supervised corpora, were engaged to manually digitize the notes (wherever direct transcripts were not available) and annotate the newly added 837 multilingual sentences with fine-grained emotion labels and then perform the causal span marking task at sentence-level. Each sentence is tagged with at most 3 emotions[6] from a set of 15 fine-grained emotion labels as described in [8]. A Fleiss-Kappa [18] score of 0.65 was attained among the three annotators, which is 0.06 points better than the work in [8] signifying that the annotations are of significantly good quality.

For each emotion (E) annotated sentence (S) in CARES_CEASE-v2.0[7], annotators were instructed to extract the causal span, *C(S)*, that adequately represented the source of the emotion *E*. We consider the span-level aggregation method discussed in [9] to mark the final causal span for a sentence *S*. If there was no specific *C(S)* for *E* in *S*, the annotators labelled the sentence as 'no cause'. Based on prior studies on span extraction, we use the macro-F1 [15] metric to measure the inter-rater agreement and attain 0.8294 F1-score, which depicts that the annotations are of substantially good quality. Our annotations process differs from [14] primarily because suicide notes are not set in a conversational setting, and thus no conversational context is available.

A sample of annotated sample is shown below:
Sentence 1: "Mom, Dad, I could not become your good son, forgive me."
Emotion: *forgiveness*; Cause: *I could not become your good son*
Sentence 2 *(Transliterated Hindi)*: "ab jine ki ichaa nhi ho rhi hai."
(English Translation): "no longer want to live."
Emotion: *hopelessness*; Cause: *no cause*

Corpus statistics: The average sentence length is 13.31 words. The longest note has 80 sentences totaling 1467 words and the shortest one contains 13 words. Table 1 shows some data about the newly collected non-English suicide notes.

Table 1. Distribution of the collected notes across various attributes. NA: Not available

Gender		**Marital status**		**Age**		**Data type**	
category	count	category	count	interval	count	category	type
				10-20	16		
Male	32	Married	15	21-30	16	Non-code-mixed	41
Female	33	Unmarried	43	31-40	3	Code-mixed	22
NA	2	NA	9	50-60	2	Transliterated	4
				NA	30		

4 Cause Recognition for Emotion in Suicide notes

We address the task of cause recognition for emotion in suicide notes as two independent sub-tasks: (A) *Emotion Cause Identification* (whether cause is present or not), (B) *Emotion Cause Extraction* (extract the causal span). Figure 1 shows the overall framework for the cause recognition setup.

[6] *forgiveness, happiness_peacefulness, love, pride, hopefulness, thankfulness, blame, anger, fear, abuse, sorrow, hopelessness, guilt, information, instructions*
[7] Dataset available at https://www.iitp.ac.in/ ai-nlp-ml/resources.html#CARES

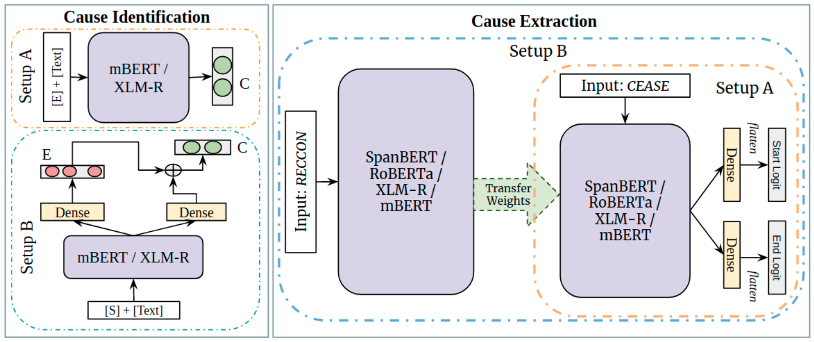

Fig. 1. Architecture of the setup for cause recognition for emotion in suicide notes.

4.1 Emotion Cause Identification

Given a sentence, s, and its associated most prominent emotion, e, the task is to identify whether s is causal or not. We propose a multitask solution (Setup B) to address the problem where the input is formed as: $< [CLS], p, [SEP], s >$, where, p indicates the associated polarity (positive, negative, neutral) of s where s is a sequence of words in the input sentence. Emotions are mapped to their associated polarity labels following the weak labelling scheme discussed in [8]. We set the max length of the input sequence as 30. The output is an emotion label from the 15-emotion tag set and a label to indicate *causal* or *non-causal*.

The input sequence is passed through a pre-trained transformer module followed by two task-specific dense layers. The softmax emotion output is fed as additional features to the task-specific causal features to enhance the output of the cause identification task, which is also generated through a softmax activation function. The model is trained using a unified loss function as shown below:

$$\lambda = \alpha * \lambda_E + \beta * \lambda_C \tag{1}$$

λ_E and λ_C are the categorical crossentropy losses for the emotion identification and cause identification task and α and β are the loss weights for the tasks, respectively. As an ablation experiment, we develop a single-task setup (Setup A) following the work in [14] where the input sequence contains the emotion label in place of polarity and outputs whether S is 'causal' or 'non-causal'.

4.2 Emotion Cause Extraction

We formulate the emotion cause extraction task for any non-neutral sentence as follows: Given a phrase s with an emotion label e, determine the causal span $c(s)$ in s that is relevant to emotion e. The input sequence (I) is formed as $< [CLS], s, [SEP], e >$. We set the max length of the input sequence as 30. We finetune four pretrained transformer-based models, Multilingual Bidirectional Encoder Representations from Transformers (mBERT) Base [5], SpanBERT Base [11], RoBERTa Base [13], and XLM-RoBERTa (XLM-R) [4] models, and for each model, the training is done as follows:

1. I is fed to the model, comprising of the sentence and emotion information.
2. Two vectors V_s and V_{cs} with dimensions equal to that of hidden states in the models are considered.

3. Each token's probability of being the start/end of the causal span is determined by a dot product between V_s/V_{cs} and the token's representation in the final layer of the model, followed by a softmax over all tokens. To compute span start and end logits, we add a dense layer on top of the hidden-states output. The sparse categorical crossentropy loss function is used in this case.
4. The model is then finetuned, allowing V_s and V_{cs} to learn along the way.

The CARES_CEASE-v2.0 dataset is used in setup A for training and testing. To train setup B, we use a transfer learning (TL) method, first training the models on the RECCON [14] dataset (only causal sentences with non-conversational context: 3613 sentences) and then finetuning on our dataset. We save the best model as per the 'Exact Match' metric [14] score on the validation set.

5 Experimental setting

The pre-trained models are sourced from the open-source libraries[8] huggingface transformers (RoBERTa, XLM-R, and SpanBERT) and tensorflowhub (mBERT). To avoid overfitting, we use ReLU activations for the dense layers (100 neurons each) and apply dropouts of 0.5 to all dense layer outputs. For the *cause identification* task, we divided the CARES CEASE-v2.0 dataset (5769 sentences) into train, test, and validation sets in the ratio 7:2:1. Only the causal utterances (1479 sentences) from the dataset are used in the *cause extraction* task. We divided the CARES CEASE-v2.0 dataset in a 7:2:1 ratio for train, test, and validation purposes. The experiments are run on an NVIDIA GeForce RTX 2080 Ti GPU and the models are optimized using Adam [12] optimizer with learning rates of 3e-5 (for mBERT and SpanBERT) and 5e-5 (for RoBERTa and XLM-R). We kept the batch size as 8 to fully utilize the GPU.

6 Results and discussion

We observe from Table 2 that for the *cause identification* task, the multitask mBERT-based system (Setup B) performs (83.20% accuracy) better than the XLM-R variant in most of the metrics. Despite being trained on highly skewed emotion data, it achieves commendable results on the emotion identification task (overall weighted-F1 of 74.48%). For the *cause extraction* task, Table 3 shows that the TL approach has proved to be effective for all the experimented models with an increase in scores for the various metrics. In the TL scenario (Setup B), the RoBERTa and XLM-R models both achieved overall top scores for three of the six metrics evaluated, with a joint highest score of 0.76 for the ROS metric. This shows the efficacy of the RoBERTa-based models compared to SpanBERT and mBERT when dealing with multi-lingual code-mixed data.

Analysis: Empirical investigation shows that learning the cause identification task together with the emotion identification task simultaneously increases performance relative to learning the task separately, regardless of the differences in pre-trained encoders. For the causal extraction task, the *Full Match (FM)* and *Partial Match (PM)* measures give a quantitative estimation of the model's performance. For quantitative evaluation, we employed an edit distance-based,

[8] https://huggingface.co/model and https://tfhub.dev/google/collections/bert/1

Table 2. Results for the cause identification task. Values in bold are the maximum scores (%) attained for a metric. A: accuracy, m-F1: macro-F1, w-F1: weighted-F1

Models	CEASE-v2.0						CARES_CEASE-v2.0					
	Setup A		Setup B				Setup A		Setup B			
	A^C	$m\text{-}F1^C$	A^C	$m\text{-}F1^C$	A^E	$w\text{-}F1^E$	A^C	$m\text{-}F1^C$	A^C	$m\text{-}F1^C$	A^E	$w\text{-}F1^E$
mBERT	81.56	78.29	82.47	80.05	70.11	70.90	81.73	80.41	**83.20**	**81.89**	75.67	**74.48**
XLM-R	80.45	79.44	80.95	79.42	70.92	60.53	80.87	79.94	81.55	79.67	**76.36**	72.81

Table 3. Results for the cause extraction task on the two setups. Values in bold are the overall maximum scores attained for a particular metric.

		CEASE-v2.0					CARES_CEASE-v2.0				
		FM (%)	*PM (%)*	*HD*	*JS*	*ROS*	*FM (%)*	*PM (%)*	*HD*	*JS*	*ROS*
Setup A	SpanBERT	23.98	24.66	0.40	0.55	0.68	31.17	17.62	0.49	0.66	0.76
	RoBERTa	34.12	21.96	0.48	0.63	0.73	28.73	19.51	0.42	0.58	0.69
	XLM-R	35.81	20.61	0.45	0.65	0.74	31.98	23.58	0.45	0.64	0.74
	mBERT	31.42	**33.11**	0.49	0.65	0.75	29.00	26.29	0.48	0.62	0.73
Setup B	SpanBERT	36.49	28.72	0.51	0.65	0.75	28.18	29.00	0.45	0.62	0.73
	RoBERTa	**38.51**	29.05	0.50	0.65	0.74	34.42	23.04	0.49	**0.67**	**0.76**
	XLM-R	34.80	26.35	0.49	0.65	**0.76**	35.23	21.41	**0.52**	0.66	**0.76**
	mBERT	33.78	26.01	0.50	0.64	0.74	29.54	28.73	0.48	0.61	0.73

token-based and sequence-based measure in the form of *Hamming distance (HD)*, *Jaccard Similarity (JS)* and *Ratcliff-Obershelp Similarity (ROS)* metrics, respectively. Manual analysis of some predicted samples with the gold annotations makes us believe that the ROS measure, based on the longest sequence matching approach, suits the training objective of causal span extraction and better estimates a model's performance from a qualitative standpoint. Although trained for span extraction tasks, we also notice that the SpanBERT model performs poorly on multilingual code-mixed data because it is exclusively trained on non-code-mixed English data. With a ROS score of 0.76, the cross-lingual XLM-R model adapts well to our multilingual data as well as the cause extraction task.

7 Conclusion and Future Work

This study focuses on addressing the task of emotion cause recognition in suicide notes by extending the size of an existing standard suicide notes dataset with multilingual data and providing gold standard emotion cause annotations on the same. Empirical results indicates that no one model can be declared to be the best at performing any of the specific sub-tasks since their performance varies with the many assessment criteria used in this study.

Future efforts might focus on extracting multiple causes from sentences and develop efficient ways to model discourse relations among the causes.

Ethical Implications

We followed the policies of using the original data and did not violate any copyright issues. The study was deemed exempt by our Institutional Review Board. The codes and data will be made available for research purposes only, after filling and signing an appropriate data compliance form.

Acknowledgement

Soumitra Ghosh acknowledges the partial support from the project titled 'Development of CDAC Digital Forensic Centre with AI based Knowledge Support Tools' supported by MeitY, Gov. of India and Gov. of Bihar (project #: P-264).

References

1. Capstick, A.: Recognition of emotional disturbance and the prevention of suicide. British Medical Journal **1**(5180), 1179 (1960)
2. Chen, Y., Hou, W., Cheng, X., Li, S.: Joint learning for emotion classification and emotion cause detection. In: Riloff, E., Chiang, D., Hockenmaier, J., Tsujii, J. (eds.) Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018. pp. 646–651. Association for Computational Linguistics (2018). https://doi.org/10.18653/v1/d18-1066, https://doi.org/10.18653/v1/d18-1066
3. Chen, Y., Hou, W., Li, S., Wu, C., Zhang, X.: End-to-end emotion-cause pair extraction with graph convolutional network. In: Proceedings of the 28th International Conference on Computational Linguistics. pp. 198–207. International Committee on Computational Linguistics, Barcelona, Spain (Online) (Dec 2020). https://doi.org/10.18653/v1/2020.coling-main.17, https://aclanthology.org/2020.coling-main.17
4. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 8440–8451. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.747, https://aclanthology.org/2020.acl-main.747
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: 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). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1423, https://aclanthology.org/N19-1423
6. Ghazi, D., Inkpen, D., Szpakowicz, S.: Detecting emotion stimuli in emotion-bearing sentences. In: Gelbukh, A.F. (ed.) Computational Linguistics and Intelligent Text Processing - 16th International Conference, CICLing 2015, Cairo, Egypt, April 14-20, 2015, Proceedings, Part II. Lecture Notes in Computer Science, vol. 9042, pp. 152–165. Springer (2015). https://doi.org/10.1007/978-3-319-18117-2_12, https://doi.org/10.1007/978-3-319-18117-2_12
7. Ghosh, S., Ekbal, A., Bhattacharyya, P.: Cease, a corpus of emotion annotated suicide notes in english. In: Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020. pp. 1618–1626. European Language Resources Association (2020), https://aclanthology.org/2020.lrec-1.201/
8. Ghosh, S., Ekbal, A., Bhattacharyya, P.: A multitask framework to detect depression, sentiment and multi-label emotion from suicide notes. Cognitive Computation pp. 1–20 (2021)
9. Gui, L., Wu, D., Xu, R., Lu, Q., Zhou, Y.: Event-driven emotion cause extraction with corpus construction. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. pp. 1639–1649. Association for Computational Linguistics, Austin, Texas (Nov 2016). https://doi.org/10.18653/v1/D16-1170, https://aclanthology.org/D16-1170

10. Ho, T., Yip, P.S., Chiu, C., Halliday, P.: Suicide notes: what do they tell us? Acta Psychiatrica Scandinavica **98**(6), 467–473 (1998)
11. Joshi, M., Chen, D., Liu, Y., Weld, D.S., Zettlemoyer, L., Levy, O.: SpanBERT: Improving pre-training by representing and predicting spans. Transactions of the Association for Computational Linguistics **8**, 64–77 (2020). https://doi.org/10.1162/tacl_a_00300, https://aclanthology.org/2020.tacl-1.5
12. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015), http://arxiv.org/abs/1412.6980
13. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized BERT pretraining approach. CoRR **abs/1907.11692** (2019), http://arxiv.org/abs/1907.11692
14. Poria, S., Majumder, N., Hazarika, D., Ghosal, D., Bhardwaj, R., Jian, S.Y.B., Hong, P., Ghosh, R., Roy, A., Chhaya, N., et al.: Recognizing emotion cause in conversations. Cognitive Computation pp. 1–16 (2021)
15. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: SQuAD: 100,000+ questions for machine comprehension of text. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. pp. 2383–2392. Association for Computational Linguistics, Austin, Texas (Nov 2016). https://doi.org/10.18653/v1/D16-1264, https://aclanthology.org/D16-1264
16. Russo, I., Caselli, T., Rubino, F., Boldrini, E., Martínez-Barco, P.: EMOCause: An easy-adaptable approach to extract emotion cause contexts. In: Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011). pp. 153–160. Association for Computational Linguistics, Portland, Oregon (Jun 2011), https://aclanthology.org/W11-1720
17. Shneidman, E.S., Farberow, N.L.: A socio-psychological investigation of suicide. In: Perspectives in personality research, pp. 270–293. Springer (1960)
18. Spitzer, R.L., Cohen, J., Fleiss, J.L., Endicott, J.: Quantification of agreement in psychiatric diagnosis: A new approach. Archives of General Psychiatry **17**(1), 83–87 (1967)
19. Talmy, L.: Toward a cognitive semantics, vol. 2. MIT press (2000)
20. Wagner, F.: Suicide notes. Danish Medical Journal **7**, 62–64 (1960)
21. Xia, R., Ding, Z.: 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. pp. 1003–1012. Association for Computational Linguistics, Florence, Italy (Jul 2019). https://doi.org/10.18653/v1/P19-1096, https://aclanthology.org/P19-1096