All Models are Wrong, But Some are Deadly: Inconsistencies in Emotion Detection in Suicide-related Tweets

Anonymous ACL submission

Abstract

Recent work in psychology has shown that people who experience mental health challenges (e.g., suicidal ideation) are more likely to express their thoughts, emotions, and feelings on social media than seeking for professional help. Distinguishing suicide-related content, such as suicide mentioned in a humorous context, from genuine expressions of suicidal ideation is essential to better understanding context and risk. In this paper, we give a first insight and analysis into the differences between emotion labels annotated by humans and labels predicted by three fine-tuned language models (LMs) for suicide-related content. We find that (i) there is little agreement between LMs for emotion labels of suicide-related Tweets and (ii) individual LMs predict similar emotion labels for all suicide-related categories. Our findings lead us to question the credibility and usefulness of such methods in high-risk scenarios such as suicide ideation detection.

1 Introduction

Each year more than 700,000 people die by suicide worldwide, where for each suicide there are many more attempts[1] and often numbers are underestimated due to under-reporting or misclassification (Organization et al., 2021). However, the majority of affected people also deny having suicidal thoughts when asked by a mental health professional (Snowdon and Choi, 2020). Developing methods to detect suicidal ideation and to distinguish it from other types of suicide-related content could help to reduce harm. Natural language processing can aid in identifying relevant features, where Language Models (LMs) have shown remarkable performance on a variety of tasks. The widespread availability of LMs via Huggingface[2] has enabled researchers to make quick emotion and sentiment predictions. One drawback of such an approach is that there is no 'quality check' to ensure that emotion and sentiment labels are correct. This may be specifically dangerous in high-stakes scenarios such as suicide ideation detection.

In this paper, we examine the results of three LMs that are fine-tuned to predict emotion labels and draw comparisons to expert human emotion annotations.

2 Related Work

Detecting suicide-related language and emotions Detection methods for suicidal intent, ideation, or risk based on deep and machine learning have evolved significantly over the past decades, and various techniques have been employed to enhance model accuracy. Traditionally, feature engineering has been a crucial component of these methods, where features extracted from text using dictionaries play a pivotal role in training machine learning models.

To overcome these limitations researchers have incorporated human annotation to obtain more fine-grained labels, e.g., on risk-levels (O'dea et al., 2015), distinctions between worrying language and flippant references to suicide (Burnap et al., 2017), content and affect of suicide-related posts (Schoene et al., 2022), or from clinical contexts (Pestian et al., 2010). Several methods have been proposed to detect suicide intent and ideation, including feature-based models with combinations of lexical features (Coppersmith et al., 2015), and psychological and affective features (Burnap et al., 2017). Work at the intersection of sentiment analysis and suicide has looked at augmenting neural networks with emotional information for ideation detection (Sawhney et al., 2021), introduce both psychological and affective features (Burnap et al., 2017) or distinguishing suicide notes from other types of content (Schoene and Dethlefs, 2016). In (Ghosh et al., 2022), a joint learning framework has

[1] https://www.who.int/news-room/fact-sheets/detail/suicide

[2]

been proposed with an additional knowledge module and claimed to have the highest cross-validation score. (Ren et al., 2015) explored the accumulated emotional data from Blogs and examined these emotional traits that are predictive of suicidal behaviors.

LMs in suicide detection and ideation Some work has already attempted to apply language models to the task of detection of suicidal ideation. TransformerRNN (Zhang et al., 2021) was trained to detect suicide notes extracted from the Reddit platform. BERT, ALBERT, Roberta, and XLNET models have shown their superiority over traditional variations like Bi-LSTM in suicide ideation from tweets on social media (Haque et al., 2020; Kodati and Tene, 2023). In an extensive study across 25 datasets from Public Health Surveillance (PHS) tasks, the PHS-BERT has demonstrated superior performance in robust and generalization capabilities (Naseem et al., 2022). Despite progress in this domain, there has been relatively little study of the robustness and consistency of LMs as applied to suicide-related text. Our work aims to extend the existing literature to better understand what kind of variation is expected when attempting to infer emotions in suicide-related text with a model that was trained on a more general corpus.

3 Methodology

3.1 Dataset

The TWISCO dataset was first introduced by (Schoene et al., 2022) and contains *3,977* Tweets annotated for suicide-related content, emotions, and VAD labels. In Table 1 we show the type of content labels and number of tweets for each category. Each Tweet was annotated by three mental health professionals for a single emotion based on Ekman's six basic emotions theory (Ekman, 1992) and a '*Neutral*' category.

Table 1: Description of TWISCO labels

Content Label	Frequency
Facts about suicidality	131
Suicide discussed philosophically/ religiously	309
Contacts for suicide-related help-seeking	51
News report, case studies or stories	291
Humorous use	165
Content not relevant	2,497
Expressing own suicidality	443
Expressing worries about suicidality of others	90
Total	3,977

Table 2: Emotions in LMs

Emotions	TWISCO	Distil-roBERTa	Distil-bert	RoB-ERTa
Anger	✓	✓	✓	✓
Disgust	✓	✓		
Fear	✓	✓	✓	
Joy	✓	✓	✓	✓
Neutral	✓	✓		
Sadness	✓	✓	✓	✓
Surprise	✓	✓	✓	

Table 3: Distribution of Emotion labels for Human annotated and LM Predicted

Emotion	TWISCO	Distil RoBERTa	Roberta	Distil bert
Neutral	1576	207	-	-
Sadness	769	1057	2012	208
Anger	554	481	428	1354
Joy	532	251	1537	-
Surprise	226	1547	-	24
Disgust	197	376	-	-
Fear	123	58	-	1121
Total	**3,977**	**3,977**	**3,977**	**3,977**

3.2 Experimental Setup

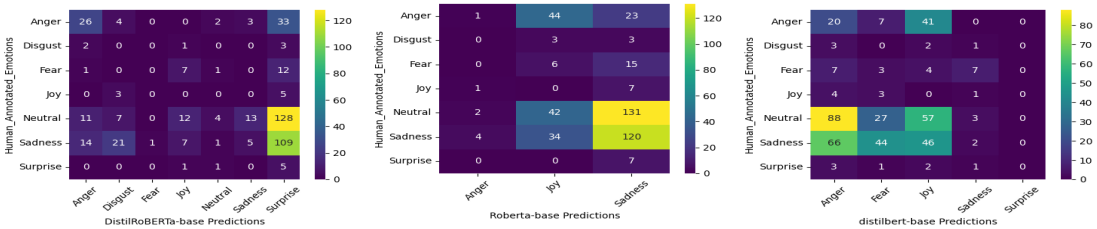
We fine-tuned three LMs to predict a single emotion label per Tweet. In Table 2 the presence of emotions in each LM is depicted. The LM proposed by Hartmann (2022) (*hereafter* Distil-RoBERTa) matches the emotion labels in TWISCO and Distilbert[3] and Twitter-roBERTa[4] only partially match. To establish a uniform approach for comparison, we have replaced the emotions '*Love*' and '*Optimism*' with '*Joy*'(for Distilbert and Roberta) following Plutchik's wheel of emotions.

4 Results

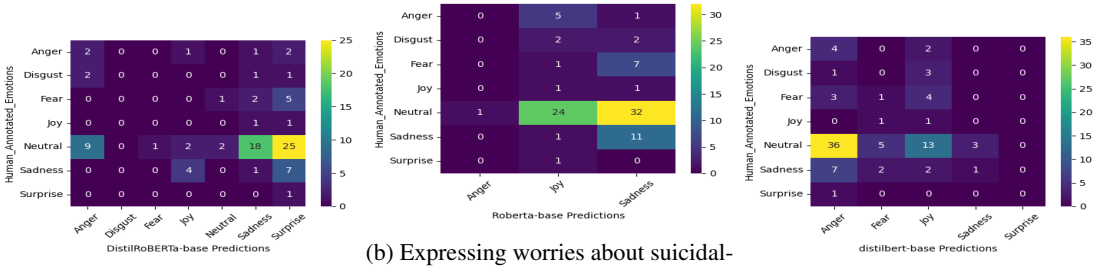
We show in Table 3 the number of annotations per emotion category across 3 LMs compared to human annotations. The emotion '*Neutral*' scores the highest based on human annotations. However, there is no agreement on the most frequent emotion across the LMs. The emotion '*Fear*' has the lowest count for both human annotation as well as Distil-RoBERTa whereas Distilbert recorded the highest count for '*Fear*'. It is observed that there are highly dissimilar patterns observed in the frequency of emotions across human annotations and the LMs employed for prediction.

[3]https://huggingface.co/bhadresh-savani/distilbert-base-uncased-emotion

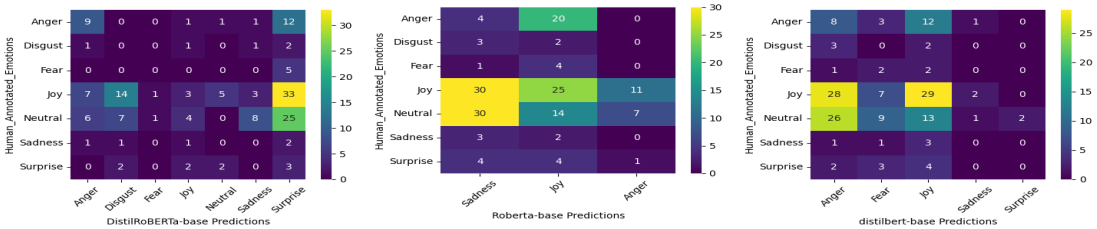
[4]https://huggingface.co/cardiffnlp/twitter-roberta-base-emotion



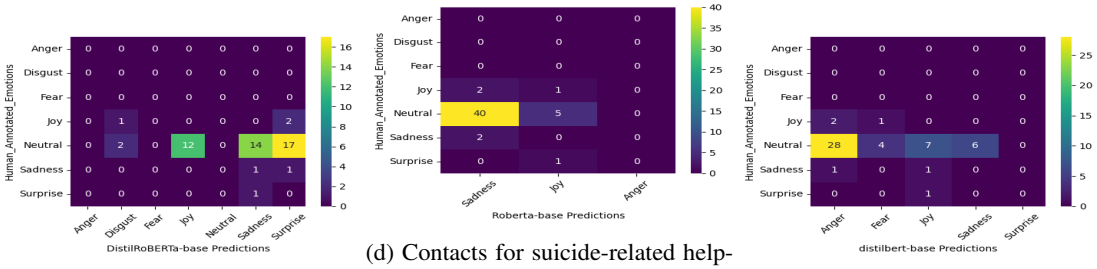
(a) Expressing own suicidality



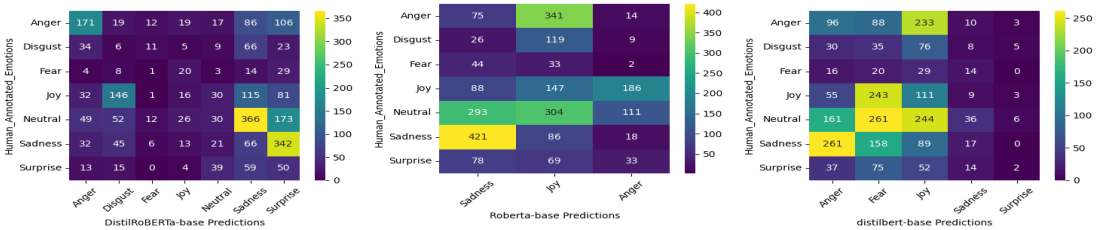
(b) Expressing worries about suicidality of others



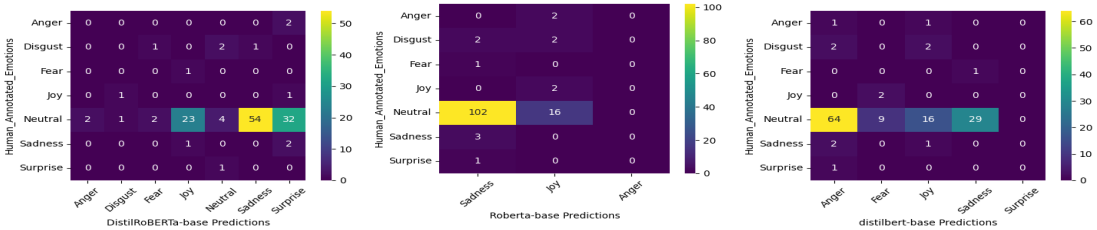
(c) Humorous use



(d) Contacts for suicide-related help-seeking

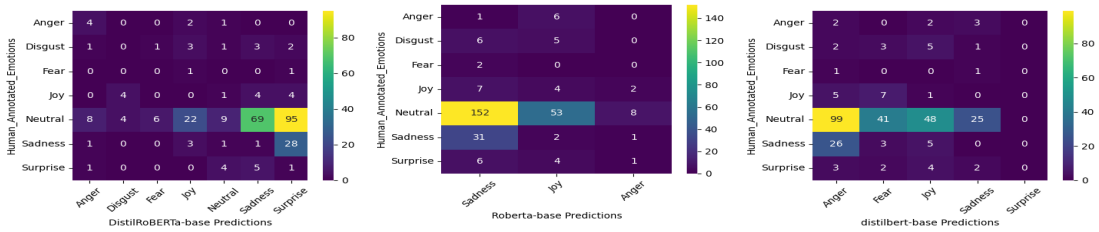


(e) Content not relevant



(f) Facts about suicidality

Figure 1: Confusion matrices for '*Expressing own suicidality*', '*Expressing worries about suicidality of others*', '*Humorous use*', '*Contacts for suicide-related help seeking*', '*Content not relevant*', and '*Facts about suicidality*' across LMs



(a) News report, case studies or stories

Figure 2: Confusion matrices for '*News reports, case studies or stories*' across LMs

To delve deeper into the performance comparison across three LMs, we plot the confusion matrices for seven categories in Figure 1 and Figure 2 and draw the following observations:

- The human annotated emotions (ground truth) exhibit variations across content categories. For instance, in Figure 1.(a) (*Expressing own suicidality*), the most prevalent emotions are '*Neutral*', '*Anger*' and '*Sadness*', conversely in Figure 1.(c) (*Humorous case*), the dominant emotions are '*Neutral*' and '*Joy*'. This variance signifies the role of content categories in determining specific emotion labels. Note that the highest dominance of emotion '*Neutral*' in human annotation is plausible (Schoene et al., 2022).

- For Distil-RoBERTa, the consistent pattern across categories indicates that the model is biased towards *Sadness*' and '*Surprise*' emotions regardless of the categories. A similar pattern for the emotions '*Sadness*' and '*Joy*' can be observed for Roberta, whereas for Distilbert, it is biased towards '*Anger*' and '*Joy*' for most of the categories.

- Among the seven categories, the *Content not relevant* appears to be the one having the most diverse range of emotions across all the three LMs. This makes sense considering the dataset scale in this category (Table 1).

- There are no consistent predicted emotions across the three LMs for any of the seven categories.

In Table 4, we compute the inter-annotater agreement between human annotations and LMs predictions using the Fleiss Kappa score (Fleiss et al., 2013). A value less than zero between human annotations and LM predictions indicates poor agreement suggesting that the observed agreement is lower than what would be expected by mere chance.

Table 4: Fleiss kappa score

LLMs	Human Annonations
DistilRoBERTa-base	-0.0878
Roberta-base	-0.0542
Distilbert-base	-0.1314

As this is our initial study, one significant limitation to point out is that the emotion labels do not align across LMs. Nevertheless, we anticipate that achieving significant performance improvement might be challenging given that Distil-RoBERTa, despite aligning with human emotions, also failed to grasp the content accurately.

5 Conclusions

In this work, we aimed to explore the variance between emotions annotated by humans and those predicted by Language Models from suicide-related tweets. We found that (i) across all three LMs there was no consistent pattern among emotions, (ii) LMs make the same predictions for minority categories that are related to suicide, and (iii) the models are biased towards certain emotions in most of the categories. This enforces the shortcomings of LMs in mirroring the human cognitive abilities in comprehending the context of tweets. This calls for the necessity for a 'quality check' when using AI-powered solutions in sensitive domains such as mental health.

We foresee many future directions for this study. To uncover the rationale behind the variations in distributions observed across the models, incorporating explainability across various categories and models would be a potential way to comprehend the emotion distribution disparities. Furthermore, providing external guidance to make LMs aware of the context of tweets would be an interesting dimension to explore.

6 Ethical considerations

There are many considerations when engaging with automated suicide-related language detection, which can relate but are not limited to (i) concerns related to linguistic aspects (e.g., linguistic imbalances and misrepresentation) and (ii) concerns related to developing, designing, and deploying datasets, language models and new algorithms to the public (e.g., issues of autonomy, justice, and harms), especially given their usefulness to build automated tools for suicide detection.

Moreover, the generalization of the results of these models/methods can lead to potential biases or false assumptions on other datasets. Therefore, it is crucial to consider the context of this work when considering to use it in similar applications.

Another important factor lies in ensuring the privacy and confidentiality of people sharing sensitive information online, adhering to consent and data policies, and avoiding potential harm or negative impacts on vulnerable individuals.

Finally, we raise the concern that the ethical guidance available to researchers working at the unique intersection of social media, psychology, linguistics, and machine learning is very limited. This is important given the increased attention from the research community on using Machine and Deep Learning in the mental health domain and suicide ideation detection.

Acknowledgements

References

Pete Burnap, Gualtiero Colombo, Rosie Amery, Andrei Hodorog, and Jonathan Scourfield. 2017. Multi-class machine classification of suicide-related communication on twitter. *Online social networks and media*, 2:32–44.

Glen Coppersmith, Ryan Leary, Eric Whyne, and Tony Wood. 2015. Quantifying suicidal ideation via language usage on social media. In *Joint statistics meetings proceedings, statistical computing section, JSM*, volume 110.

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

Joseph L Fleiss, Bruce Levin, and Myunghee Cho Paik. 2013. *Statistical methods for rates and proportions*. john wiley & sons.

Soumitra Ghosh, Asif Ekbal, and Pushpak Bhattacharyya. 2022. A multitask framework to detect depression, sentiment and multi-label emotion from suicide notes. *Cognitive Computation*, pages 1–20.

Farsheed Haque, Ragib Un Nur, Shaeekh Al Jahan, Zarar Mahmud, and Faisal Muhammad Shah. 2020. A transformer based approach to detect suicidal ideation using pre-trained language models. In *2020 23rd international conference on computer and information technology (ICCIT)*, pages 1–5. IEEE.

Jochen Hartmann. 2022. Emotion english distilroberta-base. https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/.

Dheeraj Kodati and Ramakrishnudu Tene. 2023. Identifying suicidal emotions on social media through transformer-based deep learning. *Applied Intelligence*, 53(10):11885–11917.

Usman Naseem, Byoung Chan Lee, Matloob Khushi, Jinman Kim, and Adam G Dunn. 2022. Benchmarking for public health surveillance tasks on social media with a domain-specific pretrained language model. *arXiv preprint arXiv:2204.04521*.

Bridianne O'dea, Stephen Wan, Philip J Batterham, Alison L Calear, Cecile Paris, and Helen Christensen. 2015. Detecting suicidality on twitter. *Internet Interventions*, 2(2):183–188.

World Health Organization et al. 2021. Suicide worldwide in 2019: global health estimates.

John Pestian, Henry Nasrallah, Pawel Matykiewicz, Aurora Bennett, and Antoon Leenaars. 2010. Suicide note classification using natural language processing: A content analysis. *Biomedical informatics insights*, 3:BII–S4706.

Fuji Ren, Xin Kang, and Changqin Quan. 2015. Examining accumulated emotional traits in suicide blogs with an emotion topic model. *IEEE journal of biomedical and health informatics*, 20(5):1384–1396.

Ramit Sawhney, Harshit Joshi, Lucie Flek, and Rajiv Ratn Shah. 2021. PHASE: Learning emotional phase-aware representations for suicide ideation detection on social media. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2415–2428, Online. Association for Computational Linguistics.

Annika M Schoene, Lana Bojanic, Minh-Quoc Nghiem, Isabelle M Hunt, and Sophia Ananiadou. 2022. Classifying suicide-related content and emotions on twitter using graph convolutional neural networks. *IEEE Transactions on Affective Computing*, (01):1–12.

Annika Marie Schoene and Nina Dethlefs. 2016. Automatic identification of suicide notes from linguistic and sentiment features. In *Proceedings of the 10th SIGHUM workshop on language technology for cultural heritage, social sciences, and humanities*, pages 128–133.

John Snowdon and Namkee G Choi. 2020. Undercounting of suicides: where suicide data lie hidden. *Global public health*, 15(12):1894–1901.

Tianlin Zhang, Annika M Schoene, and Sophia Ananiadou. 2021. Automatic identification of suicide notes with a transformer-based deep learning model. *Internet interventions*, 25:100422.

A Appendix

A.1 Settings

We fine-tuned the LMs for the TWISCO dataset with a batch size of 32 and a learning rate of $1e-5$ for ten epochs. Also, we have divided the dataset for a train-test split of 80 and 20 respectively.

A.2 Additional Results

The frequency of emotions within content categories is illustrated in Figure 3, depicting both human-annotated and LM-predicted emotions.

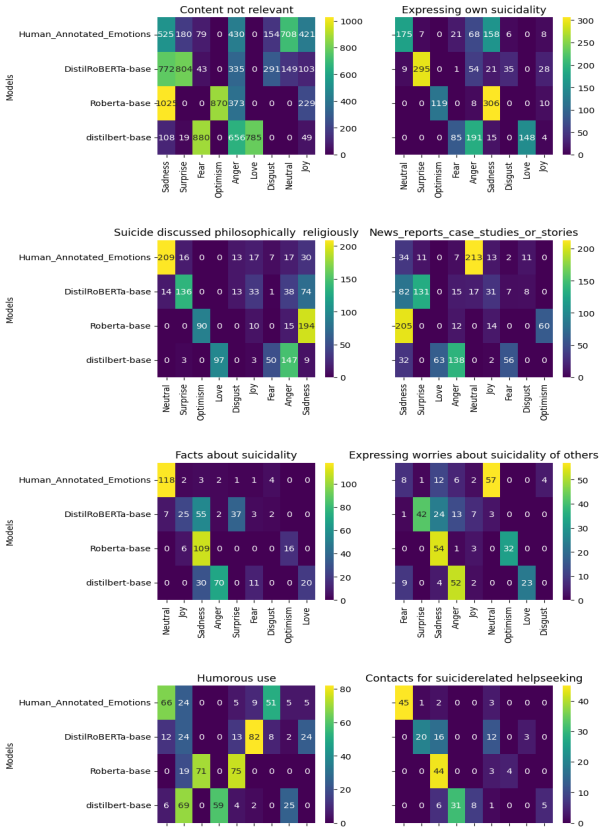


Figure 3: Distribution of emotions across categories in human annotations and LM predictions