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

Annika M. Schoene[1, 2], Resmi Ramachandranpillai[1,2], Tomo Lazovich[3], Ricardo A. Baeza-Yates[1,2]

[1]Institute for Experiential AI, [2] Northeastern University, [3] U.S. Census Bureau

Correspondence: amschoene@gmail.com

Abstract

Recent work in psychology has shown that people who experience mental health challenges are more likely to express their thoughts, emotions, and feelings on social media than share it with a clinical professional. 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 and humans 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 over 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[2]. However, the majority of affected people also deny having suicidal thoughts when asked by a mental health professional (Snowdon and Choi, 2020). In recent years, there has been tremendous growth in using Natural Language Processing (NLP) to not just identify but also understand suicidal behavior.

Many works have looked at developing methods to detect suicidal ideation with varying degrees of success and applicability to the real-world. NLP methods have been utilized to identify relevant features and more recently Language Models (LMs) have shown remarkable performance on a variety of tasks. The widespread availability of LMs via Huggingface[3] has also enabled researchers to make quick emotion and sentiment predictions. Using information available on social media, sentiment analysis has been used to detect early signs of suicidal ideation and prevent suicide attempts (). One drawback of such an approach in suicide ideation is that there is no 'quality check' to ensure that emotion and sentiment labels are correct. This may be specifically dangerous in critical applications such as suicide ideation detection.

Research Gap: State-of-the-art methods in suicidal ideation from social media currently focus on binary classification tasks, categorizing posts as either positive or negative sentiment, without considering real-time application scenarios. However, tweets often contain a spectrum of emotions tailored to specific contexts, and the absence of such nuanced analysis can affect the model's ability to identify context, leading to higher rates of false positives and false negatives. There is still a gap in understanding the consistency and robustness of these models in inferring emotions from suicide-related text. This shows a need for comprehensive examination of diverse sources of tweets related to suicide, a multifaceted approach, and the model's ability to identify embedded emotions in tweets. Moreover, this also demands a quality check on the state-of-the-art models which can only identify the presence or absence of suicide-related words or binary sentiments rather than capturing the emotions based on contexts.

Contributions: In this paper, we examine the results of three LMs that are fine-tuned to predict emotion labels from suicide-related tweets from diverse contexts and draw comparisons to human expert's emotion annotations. Our main contribu-

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

[2] https://www.who.int/data/gho/data/themes/mental-health/suicide-rates

[3] https://huggingface.co/

Proceedings of the Third Workshop on NLP for Positive Impact, pages 113–122
November 15, 2024

tions can be summarized as:

1. We show that there is no clear agreement between human annotations and LM predictions as evidenced by Inter Annotator Agreement (IAA) score (Fleiss et al., 2013) (Section 4.1; Table 4).

2. We show that LMs struggle with understanding context-dependent language, particularly in detecting humorous context and subtle expressions of distress. This can lead to misinterpretations of text and inaccurate assessments of risks (Sections 4.1 and 4.2).

3. We use both ecosystem analysis [4] (Toups et al., 2024) and Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2001) to breakdown the emotional nuances in linguistic and psychological dimensions between human annotations and LM predictions, particularly regarding emotional tone and cognitive level in multi-contextual settings (Section 4.2).

After exploring the emotional nuances between human annotations and LM predictions in identifying emotions in tweets that contain suicide-related content, our study brings critical insights with broader implications for NLP for Mental Health:

- **Gaps in contextual understanding:** This analysis unveils the lack of the LM's ability to understand contextual knowledge even after fine-tuning. Adapting ML models to contextual variations is crucial for improving the accuracy and relevance of machine learning applications in healthcare.

- **Methodological scrutiny and human-in-the-loop evaluation:** Our study involves healthcare experts in evaluating the language model's performance. This underscores the importance of methodological scrutiny and ongoing rigorous human-in-the-loop assessment of algorithms. The decisions coupled with human oversight, ensure the reliability and ethical soundness of ML-powered solutions in healthcare.

- **Psychological and linguistic analysis:** In addition to the traditional performance metrics (e.g., confusion matrix), we conduct an in-depth breakdown in terms of emotions, cognitive processes, and psychological processes. This facilitates a more nuanced understanding of the human psychological emotions embodied in the text.

- **Ecosystem analysis:** We identify tweets that are misclassified by all the LMs in the study. This is termed systemic failure (Toups et al., 2024), where certain tweets are consistently misclassified. Our analysis reveals that LMs exhibit bias toward certain parts of the tweet that contain contradictions and uncertainty, without fully capturing the emotions expressed as a whole. We advocate for ecosystem analysis to identify systemic failure when ML solutions are implemented in mental health applications.

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 been proposed with an additional knowledge mod-

[4] certain instances are always misclassified by all the models

ule 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 LMs to the task of detection of suicidal ideation. Transformer-RNN (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 in understanding what kind of variation is expected when attempting to infer emotions from multifaceted suicide-related text with a model that was trained on a more general dataset.

3 Methods

In this section, we detail the dataset description and composition, annotation process, and annotation categories present in the tweets.

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 (Valence, Arousal, and Dominance) labels. In this work, we will utulize both emotion and content labels and in Tables 1 and 2 we show the type of content and emotion labels respectively, alongside the number of tweets for each category. TWISCO was pre-processed adhering to the standard procedure of preprocessing using Ekphrasis (Baziotis et al., 2017), which involved removing user identifying details such as usernames and URLs. In compliance with Twitter's regulations, only Tweet IDs are retained for this dataset, ensuring anonymity.

3.2 Annotation Categories

The TWISCO was dataset annotated for two overarching categories, content and emotion labels.

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

Table 1: Description of TWISCO labels

Content Annotations:

- *Facts about suicidality:* These are tweets about expressing or sharing facts about suicide. While factual details regarding suicide may appear unbiased, there is research suggesting that this could potentially be vulnerable to individuals who are researching methods online.

- *Suicide discussed philosophically or religiously:* Tweets about suicide from philosophical or religious directions involve judgment and can perpetuate the stigma of suicidal thoughts, potentially deterring individuals from seeking help.

- *Contacts for suicide-related help-seeking:* Certain Tweets/posts offer guidance on where people can seek assistance and include links to resources for support.

- *News reports, case studies, or stories:* Tweets/posts discussing suicide within the context of news reports, case studies, or personal stories are considered here.

- *Humorous use:* These are tweets containing phrases associated with suicidality in a sarcastic and/or joking way that can lead to suicide intent detection algorithms prone to false positive rates.

- *Content not relevant:* Due to the data collection and curation process of TWISCO, some content is not relevant to suicide intent detection and therefore should be flagged as such.

- *Expressing own suicidality:* These are tweets from users who express their own suicidality and are experiencing a high level of distress. Identifying these users via an algorithm and assisting them could prove beneficial.

- *Expressing worries about the suicidality of others:* These are posts that express distress similar to users who are expressing their own suicidality

Emotion Annotations: Each tweet is annotated for one of Ekman's six basic emotions. An additional category called 'Neutral', has been added for instances where annotators are unsure or the emotional content of the Tweet is not apparent. One limitation of this annotation process is that a single Tweet could potentially express more than one emotion. In this case, when there was no agreement among annotators for emotion, majority voting or additional annotation rounds were instructed.

4 Experimental Results and Discussions

In this section, we provide a detailed description of the LMs employed in our study, along with their predictions and comparisons with human annotations. Additionally, we also extend our discussion of tweets that are misclassified by all the LMs in Ecosystem analysis.

Choice of Language Models: Language models including BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), and DistilBERT (Sanh et al., 2019), have been proven effective in detecting suicidal ideation from social media platforms like Twitter (Haque et al., 2020; Kodati and Tene, 2023). These models have shown superiority over traditional RNN-based methods and have proven robust performance in health surveillance tasks from Tweets. The goal of this study is to understand the ability of models that were trained on a more general corpus when attempting to infer emotions from suicide-related texts. Therefore, we have chosen three LMs: *DistilBERT*[5], *Twitter RoBERTa*[6], and *DistilroBERTa* (Hartmann, 2022), because they contain the closest matching emotion labels, are most frequently downloaded, and have been trained on similar data (e.g., Tweets), and fine-tuned on similar applications.

We fine-tune each LM to predict a single emotion label per tweet. In Table 2 the presence of emotion label for each LM is shown. The LM proposed by Hartmann (2022) called *DistilroBERTa* matches the emotion labels in TWISCO, whereas *DistilBERT* and *Twitter-RoBERTa* only partially match. To establish a uniform approach for comparison, we have replaced the emotions '*Love*' and '*Optimism*' with '*Joy*'(for *Distil-BERT* and *Twitter-RoBERTa*) following Plutchik's wheel of emotions (Plutchik, 1980).

Emotions	TWISCO	Distil-RoBERTa	Distil BERT	Twitter RoBERTa
Anger	✓	✓	✓	✓
Disgust	✓	✓		
Fear	✓	✓	✓	
Joy	✓	✓	✓	✓
Neutral	✓	✓		
Sadness	✓	✓	✓	✓
Surprise	✓	✓	✓	

Table 2: Emotion labels present in TWISCO and LMs.

Emotion	TWISCO	Distil RoBERTa	Distil BERT	Twitter RoBERTa
Neutral	1576	207	-	-
Sadness	769	1057	2082	201
Anger	554	481	1354	428
Joy	532	251	-	1537
Surprise	226	1547	24	-
Disgust	197	376	-	-
Fear	123	58	1121	-
Total	**3,977**	**3,977**	**3,977**	**3,977**

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

4.1 Comparison of Language Model Predictions

We show in Table 3 the number of annotations per emotion category across three LMs compared to human annotations in TWISCO. The label '*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*'. We observe that there are highly dissimilar patterns in the frequency of emotions across human annotations and the LMs employed for prediction.

To delve deeper into the performance comparison across three LMs, we plot the confusion matrices for all content-related categories (as shown in Table 1), where in Figures 1 - 8 we show confusion matrices for each LM compared to TWISCO's human annotations.

- The human annotated emotions (ground truth) reflect the contextual variations. For instance, in Figure 1, the most prevalent emotions in human annotation are '*Neutral*', '*Anger*' and

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

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

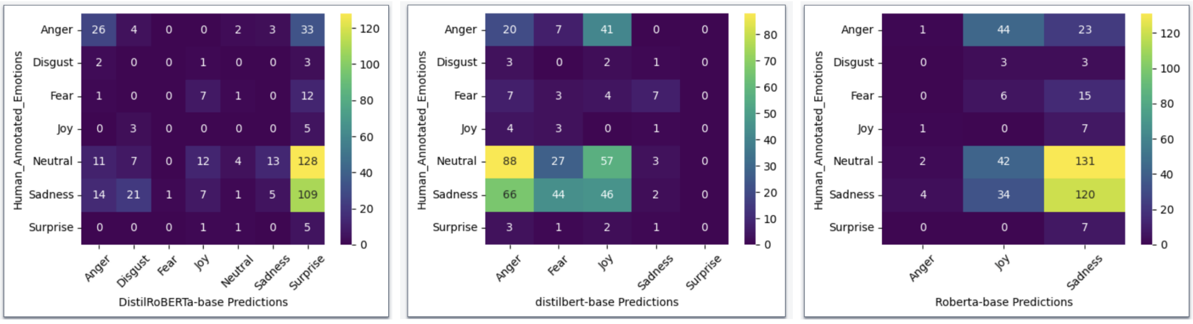


Figure 1: Confusion Matrices for the content label ‘Expressing own suicidality’

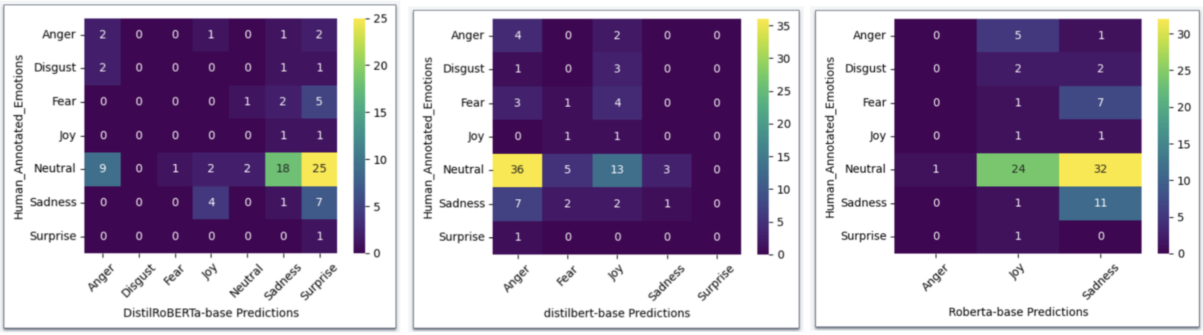


Figure 2: Confusion Matrices for the content label ‘Expressing worries about suicidality of others’

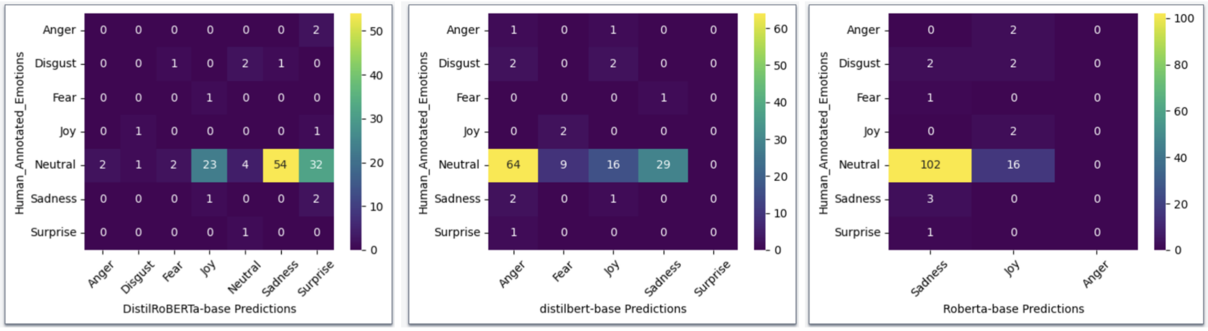


Figure 3: Confusion Matrices for the content label ‘Facts about suicidality’

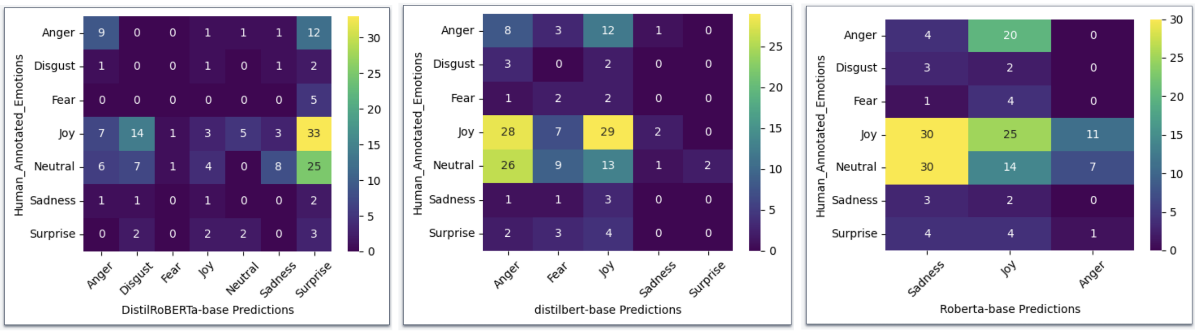
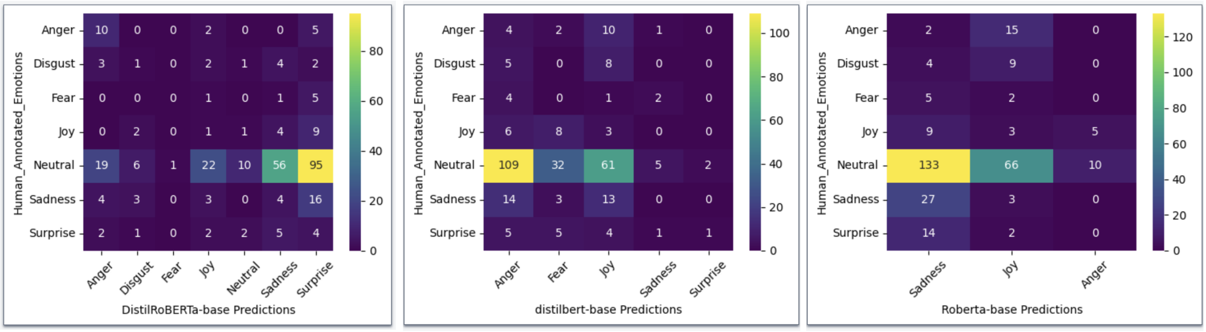


Figure 4: Confusion Matrices for the content label ‘Humorous use’

’*Sadness*’, conversely in Figure 4 (Label: *Humorous use*), the dominant emotions are ’*Neutral*’ and ’*Joy*’. This variance signifies the role of content categories in determining specific emotion labels. Note that the ’*Neutral*’ is the most frequent human annotation, which is plausible as (i) the majority of tweets do not express suicidal ideation or content, (ii) some

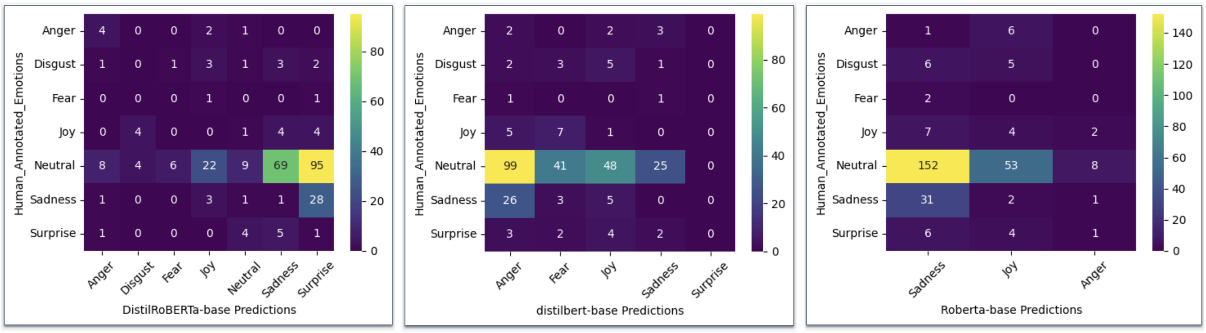


	Anger	Disgust	Fear	Joy	Neutral	Sadness	Surprise
Anger	10	0	0	2	0	0	5
Disgust	3	1	0	2	1	4	2
Fear	0	0	0	1	0	1	5
Joy	0	2	0	1	1	4	9
Neutral	19	6	1	22	10	56	95
Sadness	4	3	0	3	0	4	16
Surprise	2	1	0	2	2	5	4

	Anger	Fear	Joy	Sadness	Surprise
Anger	4	2	10	1	0
Disgust	5	0	8	0	0
Fear	4	0	1	2	0
Joy	6	8	3	0	0
Neutral	109	32	61	5	2
Sadness	14	3	13	0	0
Surprise	5	5	4	1	1

	Sadness	Joy	Anger
Anger	2	15	0
Disgust	4	9	0
Fear	5	2	0
Joy	9	3	5
Neutral	133	66	10
Sadness	27	3	0
Surprise	14	2	0

Figure 5: Confusion Matrices for the content label ‘Suicide discussed philosophically or religiously’

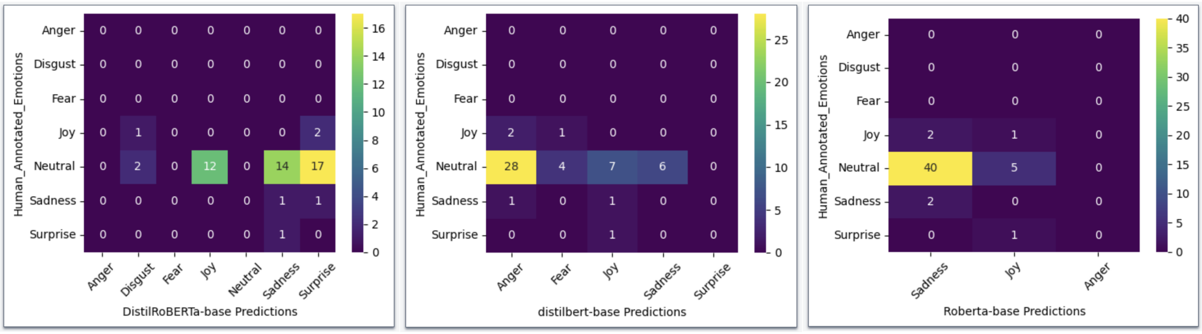


	Anger	Disgust	Fear	Joy	Neutral	Sadness	Surprise
Anger	4	0	0	2	1	0	0
Disgust	1	0	1	3	1	3	2
Fear	0	0	0	1	0	0	1
Joy	0	4	0	0	1	4	4
Neutral	8	4	6	22	9	69	95
Sadness	1	0	0	3	1	1	28
Surprise	1	0	0	0	4	5	1

	Anger	Fear	Joy	Sadness	Surprise
Anger	2	0	2	3	0
Disgust	2	3	5	1	0
Fear	1	0	0	1	0
Joy	5	7	1	0	0
Neutral	99	41	48	25	0
Sadness	26	3	5	0	0
Surprise	3	2	4	2	0

	Sadness	Joy	Anger
Anger	1	6	0
Disgust	6	5	0
Fear	2	0	0
Joy	7	4	2
Neutral	152	53	8
Sadness	31	2	1
Surprise	6	4	1

Figure 6: Confusion Matrices for the content label ‘News reports, case studies or stories’

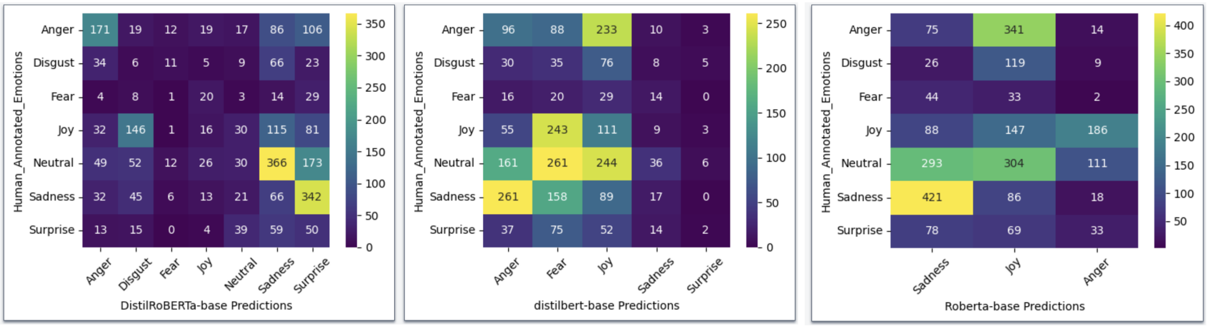


	Anger	Disgust	Fear	Joy	Neutral	Sadness	Surprise
Anger	0	0	0	0	0	0	0
Disgust	0	0	0	0	0	0	0
Fear	0	0	0	0	0	0	0
Joy	0	1	0	0	0	0	2
Neutral	0	2	0	12	0	14	17
Sadness	0	0	0	0	0	1	1
Surprise	0	0	0	0	0	1	0

	Anger	Fear	Joy	Sadness	Surprise
Anger	0	0	0	0	0
Disgust	0	0	0	0	0
Fear	0	0	0	0	0
Joy	2	1	0	0	0
Neutral	28	4	7	6	0
Sadness	1	0	1	0	0
Surprise	0	0	1	0	0

	Sadness	Joy	Anger
Anger	0	0	0
Disgust	0	0	0
Fear	0	0	0
Joy	2	1	0
Neutral	40	5	0
Sadness	2	0	0
Surprise	0	1	0

Figure 7: Confusion Matrices for the content label ‘Contacts for suicide-related help-seeking’



	Anger	Disgust	Fear	Joy	Neutral	Sadness	Surprise
Anger	171	19	12	19	17	86	106
Disgust	34	6	11	5	9	66	23
Fear	4	8	1	20	3	14	29
Joy	32	146	1	16	30	115	81
Neutral	49	52	12	26	30	366	173
Sadness	32	45	6	13	21	66	342
Surprise	13	15	0	4	39	59	50

	Anger	Fear	Joy	Sadness	Surprise
Anger	96	88	233	10	3
Disgust	30	35	76	8	5
Fear	16	20	29	14	0
Joy	55	243	111	9	3
Neutral	161	261	244	36	6
Sadness	261	158	89	17	0
Surprise	37	75	52	14	2

	Sadness	Joy	Anger
Anger	75	341	14
Disgust	26	119	9
Fear	44	33	2
Joy	88	147	186
Neutral	293	304	111
Sadness	421	86	18
Surprise	78	69	33

Figure 8: Confusion Matrices for the content label ‘Content not relevant’

content (e.g.: news) may not evoke an emotional reaction in the reader, or (iii) it might not be clear what the emotional content is (Schoene et al., 2022).

- Predictions by *Distil-roBERTa* lack the contextual variations and show consistent patterns across categories indicating that the model is biased towards *Sadness*’ and ’*Surprise*’ emotions regardless of the content categories. A similar pattern for the emotions '*Sadness*' and '*Joy*' can be observed for *Twitter-RoBERTa*, whereas for *DistilBERT*, it is biased towards '*Anger*' and '*Joy*' for most of the categories.

- There are (i) no consistent predicted emotions across the three LMs for any of the seven suicide-related content categories and (ii) no agreement between the human annotations and those predicted by LMs in any of the seven categories.

Comparing In Table 4, we compute the IAA score 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.

LLMs	Human Annonations
DistilRoBERTa-base	-0.0878
Twitter-Roberta	-0.0542
DistilBERT-base	-0.1314

Table 4: Fleiss kappa scores for each LM compared to the ground truth labeled provided in TWISCO .

4.2 Qualitative Ecosystem Analysis

Ecosystem analysis was first introduced in (Toups et al., 2024), where instead of examining a single model, an array of models were deployed for a specific context (e.g.: predicting if a candidate is hired or not) and subsequently analyzed for performance. This type of analysis can be useful in identifying systemic failure, where for our use case some instances are always misclassified by all selected LMs. Here, we identified instances prone to systemic failure where tweets were misclassified by all three models when compared to their respective human annotations in TWISCO. To clearly understand the rationale behind these misclassified instances, we used LIWC (Pennebaker et al., 2001) for our anallysis on those instances (see Figures 9 and 10).

Example 1: Consider a Tweet that reads '*i wanna die more than anything i could ever create from this earth...*' from the content category *Expressing own suicidality*. The human emotion annotation for this tweet is *Sadness* while the emotions predicted by *DistilroBERTa*, *Twitter-roBERTa*, and *DistilBERT* are *Surprise*, *Surprise*, and *Joy* respectively. We broke down the LIWC categories in terms of Emotional Analysis and Psychological Processes to gain further insight into the correlation between emotions and cognitive thinking (see Figure 9). The dominant emotion annotated in TWISCO matches with the LIWC emotion categories where a combination of *Sadness* and *Negative* identifies *Sadness* regardless of the presence of *Excited* as the overarching sentiment is *Negative*.

The *Analytic* dimension in the Psychological Processes category reflects the degree of analytical thinking and cognitive complexity present in the text. High scores indicate logical and rational thinking, while low scores suggest a less analytical style. A high value for *Tone* indicates the intensity of emotion present in the sentence. A high value in the combination of *Tentativeness* and *Differences* (called as *tentat*, and *differ* in Figure 9) indicates inconsistency or unsure expressed in the Tweet. This indicates a part of the Tweet that contains contradictions, uncertainty, or inconsistencies. There is also a high value for the *Discrepancy* dimension (called *discrep*), which indicates a difference between the current state a person is in and a more complete state they would like to be in (Boyd et al., 2022).

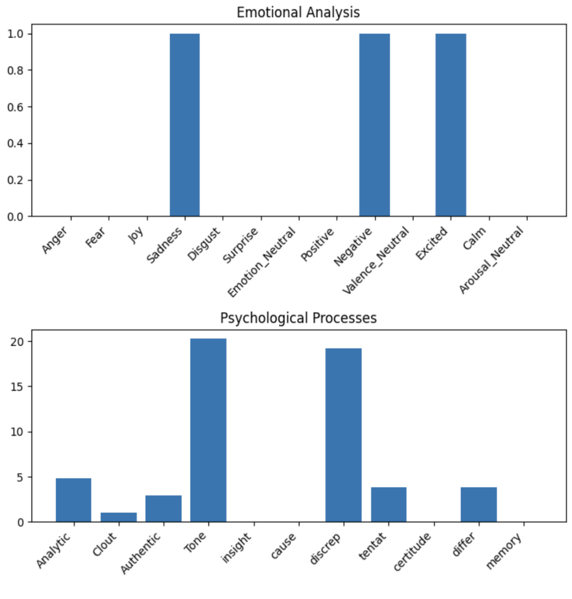


Figure 9: LIWC analysis of Example 1 using a Tweet from the *Expressing Own Suicidality* content category.

Example 2: Here we consider a tweet from the *Humorous use* content category that reads '*its like wanting to commit suicide w out actually dying*'. Again, we analyze the tweet using LIWC (see Figure 10) for both Emotions and Psychological Processes. The human annotation for this tweet is *Joy* whereas the LMs, namely *DistilroBERTa*, *Twitter-roBERTa*, and *DistilBERT* predicted the emotion labels as *Surprise*, *Sadness*, and *Anger* respectively. LIWC identifies *Joy* and *Positive* as the dominant emotion categories in addition to a high score for *Arousal Neutral*, which might indicate that the emotional content of the text does not evoke strong feelings.

LIWC's Psychological Process categories give a high score in the *Analytic*, *Clout*, and *Tone* indicating that the cognitive element, confidence, and intensity of the sentence is clear, which in turn promotes *Joy*. Furthermore, the *Discrepancy* and *certitude* of the tweet are similar, yet a human annotator could comprehend and amalgamate the emotions and psychological dimensions. Each LM misclassified the tweet overlooking the context in which the tweet is being used, as shown by the LIWC breakdown analysis.

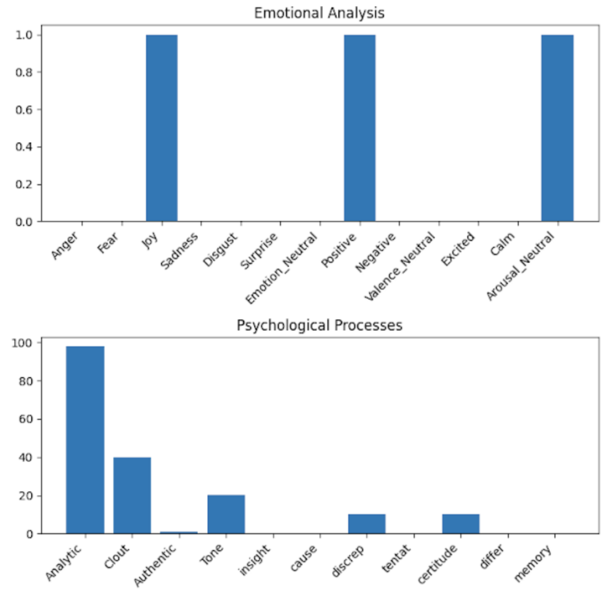


Figure 10: LIWC analysis of Example 2 using a Tweet from the *Humorous use* content category.

Overall, this shows us that (i) human annotators can encompass the consideration of all these dimensions and (ii) LIWC is more aligned with human emotion label judgments, whereas LMs might be somewhat limited in how they interpret emotional content from single sentences. One possible reason for this could be due to biases towards certain aspects of a sentence or assigning more importance to a specific word rather than considering the context beyond the company a word keeps.

5 Conclusions

In this work, we explored 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 limited consensus among models and between models and humans, (ii) LMs make the same predictions for minority categories that are related to suicide, (iii) the models are biased towards certain emotions in most of the categories, and (iv) the models cannot identify the correlation between emotions and psychological process for certain instances, that are prone to systemic failure as evidenced by LIWC breakdown. This enforces the shortcomings of LMs in mirroring the human cognitive abilities in comprehending the context of tweets and shows that there is an increased need for a 'quality check' when using AI-powered solutions in critical and sensitive application areas such as mental health.

Limitations and Future Directions This is a first study on using LMs to predict emotions in critical applications (e.g.: suicide-related content) and there are several limitations: (i) the emotion labels do not align across all LMs and with the original corpus, (ii) the dataset itself is relatively small and an analysis over other similar dataset would be beneficial to see if these initial findings generalize across datasets, and (iii) we only used a limited number of LMs and a comparison across more systems would be helpful to validate initial patterns. However, *Distil-RoBERTa* aligns fully with TWISCO's annotated emotion labels and also failed to capture the emotional content compared to human annotations. Therefore, we would like to see future research to further (i) investigate if these patterns generalizes over different datasets, (ii) include multiple other LMs into Ecosystem analysis, and (iii) conduct a more in-depth qualitative ecosystem analysis over multiple label categories. In addition to this, we would like to uncover the rationale behind the variations in distributions observed across the LMs, incorporating explainability across various categories and LMs would be a potential way to comprehend the emotion distribution disparities. Finally, 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 are related but are not limited to concerns (i) regarding to linguistic aspects (e.g., linguistic imbalances and misrepresentation), where certain phrases or words may not translate well to other cultures and languages and (ii) related to developing, designing, and deploying datasets, LMs 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 using 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.

Acknowledgments

References

Christos Baziotis, Nikos Pelekis, and Christos Doulkeridis. 2017. Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, pages 747–754.

Ryan L Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W Pennebaker. 2022. The development and psychometric properties of liwc-22. *Austin, TX: University of Texas at Austin*, 10.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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.

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. *arXiv preprint arXiv:1907.11692*.

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.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

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.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. In *Theories of emotion*, pages 3–33. Elsevier.

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.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

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.

Connor Toups, Rishi Bommasani, Kathleen Creel, Sarah Bana, Dan Jurafsky, and Percy S Liang. 2024. Ecosystem-level analysis of deployed machine learning reveals homogeneous outcomes. *Advances in Neural Information Processing Systems*, 36.

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.