Benchmarking Suicide-Related Language: A Study on Interpretability and Explainability

Anonymous ACL submission

Abstract

Individuals experiencing mental health challenges often share their thoughts and emotions on social media rather than seeking professional support, making it crucial to accurately distinguish genuine suicidal ideation from mentions of suicide in humor or figurative language. However, automated suicide detection faces challenges such as data scarcity, ambiguity in suicidal expressions, and inconsistencies between human and model predictions. To address these issues, we (i) expand an existing dataset via expert annotation, enhancing data diversity and representation, and (ii) benchmark the performance of 8 state-of-the-art language models (LMs), including both general-purpose and domain-specific models, and (iii) conduct a category-wise performance analysis to evaluate their effectiveness in detecting suicide-related content. Our findings demonstrate that domain-specific models, particularly MentalRoBERTa and MentalBERT, outperform general-purpose models, especially as dataset size increases. To gain further insight into LM behaviour we perform interpretability and explainability analyses, examining token importance scores to identify misclassification patterns. Results indicate that models over-rely on emotionally charged keywords, often misclassifying humor, figurative language, and personal distress expressions. Additionally, we conduct N-gram analysis across content categories, revealing significant linguistic overlap, which pose challenges for precise classification.

1 Introduction

Suicide remains a major public health crisis in the United States, consistently ranking among the leading causes of death. According to the Centers for Disease Control and Prevention[1], more than 48,000 people died by suicide in 2021 alone, with rates steadily increasing over the past two decades. This alarming trend highlights the urgent need for effective suicide prevention strategies, especially in digital spaces where individuals are increasingly vocal about their mental health struggles. Social media platforms, in particular, have become a critical avenue for detecting early warning signs of suicide, as individuals often share their thoughts and emotions online before seeking professional help. Natural Language Processing (NLP) has emerged as a powerful tool for identifying suicidal ideation in social media posts, offering scalable and automated methods to support mental health interventions. By analyzing textual cues, sentiment patterns, and contextual information, NLP models can help detect at-risk individuals and enable timely interventions.

In recent years, Language Models (LMs) (Koroteev, 2021) and Large Language Models (LLMs) (Naveed et al., 2023) have demonstrated remarkable capabilities in understanding and generating human language. Trained on vast amounts of textual data, these models show promise in detecting nuanced language patterns, including those indicative of suicidal ideation. However, applying LMs and LLMs to suicide detection requires careful fine-tuning and domain-specific adaptation to differentiate genuine distress from other contextual uses of suicide-related language, such as humor, awareness discussions, or figurative expressions.

For effective training and fine-tuning, high-quality, diverse, and well-annotated datasets are essential. However, data scarcity remains a significant bottleneck in suicide detection research. Existing datasets are often limited in size, biased toward specific suicide-related contexts, or lack sufficient representation of the various ways individuals discuss suicide. Many datasets focus solely on binary classification—suicidal vs. non-suicidal—overlooking the complexity of suicide-related discourse. This lack of comprehensive, multicategorical datasets hinders model performance, leading to increased false positives and false nega-

[1] https://www.cdc.gov/nchs/pressroom/nchs_press_releases/2022/20220930.htm

tives.

To address these challenges, our contributions can be summarized as follows:

- **Dataset Expansion:** We extend the existing dataset by incorporating additional tweets across multiple suicide-related categories, thereby enhancing its diversity and representativeness.
- **Benchmarking:** We benchmark eight state-of-the-art language models (LMs) on both the old and new versions of the dataset to evaluate their effectiveness in suicide content detection.
- **Interpretability:** We perform N-gram extraction for each category to identify key linguistic patterns, providing insights into recurring expressions and terminologies indicative of different categories of suicidal content.
- **Explainability:** We emphasize model explainability, utilizing techniques such as attention visualization and feature attribution to understand how the models arrive at their predictions.

In summary, our study aims to advance NLP-based suicide detection by enhancing dataset quality, refining model evaluation, and promoting a more transparent and interpretable approach to suicide content detection.

2 Related Works

NLP for Suicide Detection: The identification of suicide-related language and emotional expressions has seen significant advancements, particularly in detecting suicidal intent, ideation, and risk using deep learning and machine learning techniques. Over the years, researchers have explored various approaches to improve model accuracy. Historically, feature engineering has played a key role in these methods, with textual features extracted through dictionary-based techniques serving as essential inputs for training machine learning models.

To address these challenges, researchers have integrated human annotation to refine labeling granularity, such as assessing risk levels (O'dea et al., 2015), differentiating between concerning language and casual mentions of suicide (Burnap et al., 2017), analyzing both content and emotional tone in suicide-related posts (Schoene et al., 2022), and incorporating insights from clinical settings (Pestian et al., 2010). Various approaches have been introduced for detecting suicidal intent and ideation, including feature-based models that leverage lexical attributes (Coppersmith et al., 2015), as well as psychological and affective markers (Burnap et al., 2017).

Research at the intersection of sentiment analysis and suicide detection has focused on enhancing neural networks with emotional context to improve ideation recognition (Sawhney et al., 2021), integrating psychological and affective characteristics (Burnap et al., 2017), and distinguishing suicide notes from other textual content (Schoene and Dethlefs, 2016). Additionally, (Ghosh et al., 2022) proposed a multitask learning framework incorporating a knowledge module, achieving the highest cross-validation score. Furthermore, (Ren et al., 2015) analyzed emotional patterns in blogs, identifying traits predictive of suicidal behavior.

LMs for Suicide Detection and Ideation: Several studies have explored the application of language models (LMs) for detecting suicidal ideation. TransformerRNN (Zhang et al., 2021) was specifically trained to identify suicide notes extracted from Reddit. Additionally, models such as BERT, ALBERT, RoBERTa, and XLNet have outperformed traditional approaches like Bi-LSTM in detecting suicidal ideation from tweets and other social media posts (Haque et al., 2020; Kodati and Tene, 2023). A comprehensive evaluation across 25 Public Health Surveillance (PHS) datasets found that PHS-BERT exhibited strong performance, particularly in terms of robustness and generalization (Naseem et al., 2022). However, despite advancements in this area, there has been limited research on the consistency and reliability of LMs when applied to suicide-related text (Schoene et al., 2024). A comparative analysis of BERT-based LMs and LLMs was conducted in (Oliveira et al., 2024).

Datasets for Suicide Identification. Old TWISCO (Schoene et al., 2022). Guidelines for how to create a suicide dataset (Parsapoor et al., 2023). Suicide risk level prediction and suicide trigger detection: a benchmark dataset (Li et al., 2022). Suicide attempt and ideation events dataset (Rawat et al., 2022)

Explainability in Suicide Detection: The classification of suicidal thoughts in binary settings has been investigated in (Islam et al., 2023). Sim-

ilarly, (Bouktif et al., 2025) utilized tweets from the COVID-19 period to identify suicidal thoughts using a deep learning neural network, coupled with explainable AI (XAI) techniques to analyze feature attributions. However, both studies are limited to binary classification tasks and primarily focus on deep neural networks.

3 Dataset Expansion and Downsampling

TWISCO was originally proposed by Schoene et al. (2022) and contains 112,969 tweets, of which 3,977 tweets were annotated by psychologists for suicide-related language, emotion labels, and Valence, Arousal, and Dominance. For a full description of each label category, see Appendix A, and for the label distribution across each dataset, see Table 1.

Expanded TWISCO: To expand this dataset, we randomly selected 6,181 Tweets from the original unannotated TWISCO dataset and performed six rounds of annotation. For this, we recruited three annotators with annotation experience and training in psychology, who were then provided with annotation guidelines and performed the task on . Each annotator was assigned the same set of Tweets for labeling, with a single label required for each of the specified content category. To assess inter-annotator agreement, we employed Fleiss' Kappa score (Joseph and Fleiss, 2023), an extension of Cohen's Kappa score (McHugh, 2012) that accounts for multiple independent annotators rating categorical variables. The Fleiss Kappa score for our annotation is 0.705 indicating a substantial agreement among annotators.

Downsampled TWISCO: Given the overrepresentation of the *Content not relevant* category (see Table 1), we downsampled this catgeory by 10% to improve model generalizaiton and avoid overfitting. The downsampled TWISCO version contains a total of 3,208 tweets, ensuring a more balanced distribution and was used in our subsequent experiments.

4 Experiments

Our choice of LMs for detecting suicide content categories follows a previously established selection approach (Mohammadi et al., 2024; Schoene et al., 2023), aimed at capturing both general and domain-specific language nuances. To benchmark our expanded dataset and assess the difficulty of the task, we first used a *Maximum Entropy* classifier (Nigam et al., 1999). We then selected the following LMs for performance comparison, setting a learning rate of 0.00001 and a maximum sequence length of 64 for each model, except for *Bio-Clinical BERT*, which was set to 256:

- RoBERTa (Yinhan et al., 2019): A general-purpose model known for its robustness in understanding contextual word meanings.
- BERT (Devlin et al., 2018): A versatile model that excels in capturing contextual nuances.
- Twitter RoBERTa: It is tailored to social media language, enhancing understanding of informal text.
- MentalBERT (Ji et al., 2021): Specialized in mental health-related text for detecting expressions of suicidality and distress.
- Mental RoBERTa: Similar to MentalBERT, with a focus on mental health nuances in text.
- PsychBERT (Vajre et al., 2021): Focuses on psychological and mental health-related text for more nuanced analysis.
- Bio_ClinicalBERT: Provides a clinical perspective, useful for identifying structured or clinical terms.
- XLM RoBERTa: Multilingual model that broadens detection across diverse linguistic variations.

We use Precision, Recall, and F1 scores to evaluate the LMs as these metrics provide a comprehensive understanding of model performance.

Choosing LMs Over LLMs for Suicide Detection: Interpretability and Ethical Considerations: We opted to use LMs instead of LLMs for suicide-related content detection due to several critical factors. First, domain-specific LMs such as *MentalRoBERTa* are fine-tuned on mental health-related datasets, enabling them to better capture the nuanced and sensitive nature of suicide-related language compared to general-purpose LLMs. Another key consideration is interpretability. Suicide detection is a highly sensitive task that requires explainable model decisions, and LMs provide greater transparency through token importance analysis, making them more suitable for this application. Moreover, ethical and safety concerns arise

Suicide-related Content Label	TWISCO [13]	Expanded TWISCO	Downsampled TWISCO
Facts about suicidality	131	178	178
Suicide discussed philosophically or religiously	309	505	505
Contacts for suicide-related help-seeking	51	68	68
News reports, case studies, or stories	291	362	362
Humorous use	165	412	412
Content not relevant	2497	7722	772
Expressing own suicidality	443	784	784
Expressing worries about suicidality of others	90	127	127
Total	**3977**	**10,158**	**3208**

Table 1: Overview of the original TWISCO dataset, the expanded TWISCO dataset, and the downsampled *content not relevant* category in the Downsampled TWISCO.

with LLMs, as they may generate unintended biases or hallucinations, leading to potentially harmful misclassifications (Hua et al., 2024; Liu et al., 2023). Additionally, they pose risks related to data retention and inference leakage, where user-generated content might be inadvertently stored or used to refine future models (Yao et al., 2024). LLMs also require significantly more computational resources, memory, and energy, making them impractical in resource constraint settings (e.g.; academic research). Privacy concerns are another limitation, as LLMs trained on vast web-scale data may inadvertently store sensitive information, raising ethical and legal compliance issues in domains where GDPR, HIPAA, or other data protection regulations apply (Athanasopoulou, 2024; Riad et al., 2024). Furthermore, suicide-related language varies significantly across cultures and demographics, and while domain-specific LMs are trained on curated datasets that account for these variations, general-purpose LLMs may misinterpret culturally specific expressions of distress, leading to higher false positive or false negative rates (Schoene et al., 2025). Given these limitations we opted to only focus on domain-specific LMs in this study, but will work towards using LLMs in future research.

5 Results

In Table 2, we report the results of our experiments for TWISCO and Downsampled TWISCO respectivley. We find that *MentalRoBERTa* is the best-performing model across all LMs benefiting from specialized fine-tuning on mental health data. *MentalBERT* also performs well but lags slightly behind in recall and F1 scores. Both models outperform general-purpose models due to their domain-specific tuning. *PsychBERT* and *BioClinicalBERT* underperform, possibly because both models were trained on out-of-domain data. Although the content is related, this suggests that the type of data (e.g., social media) is more important than the topic (e.g., texts about mental health). Similarly, models trained on large amounts of social media data (e.g., *RoBERTa*, *BERT*, and *Twitter RoBERTa*) outperform models that are more context-specific. Overall, the dataset expansion led to notable improvements, especially for domain-specific models. We hypothesize that F1 scores would improve further as more data becomes available for this task; however, this is left for future research.

LM	TWISCO			Downsampled TWISCO		
	P	R	F1	P	R	F1
MaxEnt Classifier	0.39	0.44	0.37	0.31	0.39	0.30
RoBERTa	0.50	0.47	0.47	0.59	0.55	0.54
BERT	0.45	0.43	0.44	0.51	0.50	0.50
Twitter RoBERTa	0.61	0.53	0.56	0.59	0.56	0.57
MentalBERT	0.59	0.46	0.49	0.56	0.54	0.55
MentalRoBERTa	**0.65**	**0.58**	**0.57**	**0.72**	**0.67**	**0.64**
PsychBert	0.35	0.42	0.38	0.47	0.46	0.46
BioClinicalBERT	0.58	0.38	0.40	0.60	0.42	0.41
XLM RoBERTa	0.44	0.44	0.43	0.56	0.54	0.54

Table 2: Macro-averaged scores for each LM on TWISCO and Downsampled TWISCO.

Category-wise performance analysis: We report and analyze the category-wise performance metrics for each selected LM. In Appendix B, we provide a detailed overview in Figures 4 and 5 of the metrics for each content label. As noted previously, domain-specific models, particularly *MentalRoBERTa* and *MentalBERT*, consistently outperform general-purpose models across all categories. These improvements are especially pronounced in categories such as *Expressing own suicidality* and *News reports, case studies, or stories*, where *MentalRoBERTa* shows significant improvements in precision, recall, and F1 scores. The *Content not relevant* category exhibits strong performance across all models due to it being the majority class in this dataset. Among the eight categories of suicidal content, three—*Contacts for suicide-related help-seeking*, *Facts about suicidality*, and *Expressing worries about others' suicidality*—are minority categories. Notable improvements were ob-

served in these categories across most models in the downsampled (new version) TWISCO dataset compared to the older version, most likely due to the increased number of samples per category. A significant improvement was observed in the *Expressing worries about the suicidality of others* category in the downsampled TWISCO dataset using *MentalRoBERTa* only.

6 Interpretability and Explainability

With the increasing interest in explainability for high-risk applications, we focus on investigating the best-performing LM, MentalRoBERTa. We used *transformers-interpret*[2] library for extracting the attention scores of each token in a tweet.

Explainability: The visualization in Figure 1 illustrates token importance scores for *MentalRoBERTa*, highlighting incorrect predictions. The color intensity represents the importance of individual tokens in determining the predicted labels. Several misclassifications are evident, such as the label *Humorous use* being predicted as *Expressing Own Suicidality*, and *Expressing Own Suicidality* being classified as *Content not relevant*. Words with strong semantic weight, such as 'kill', 'suicidal' and 'hate', tend to have higher importance scores, indicating the model's sensitivity to specific language. However, this raises potential concerns regarding bias and misclassification, particularly in differentiating between humor, personal experiences, and news content. The errors suggest that the model may struggle with contextual understanding, relying heavily on certain keywords while neglecting broader context. This highlights challenges in distinguishing between categories and possibly points to a larger issue in using LMs for accurately predicting suicide content or assessing risk in real-world applications.

Figure 2 shows the token attention scores for correctly classified predictions. These instances exhibit distinct patterns that contribute to the model's accuracy. High importance scores are assigned to keywords closely related to the category, such as 'suicide', 'kill', 'death', 'psychiatric', and 'emergency', indicating that the model strongly relies on these terms for classification.Additionally, contextual clarity plays a significant role, as sentences that clearly align with their intended category—such as help-seeking phrases in the *Contacts for help-seeking* category are more likely to be classified correctly. Another notable trend is the balanced distribution of importance scores across multiple words, rather than a single dominant token, suggesting that the model considers the broader context in accurate classifications. Furthermore, categories such as *Facts about suicide* and *Contacts for help-seeking* show particularly strong importance scores for key terms, reinforcing their alignment with factual information or emergency assistance.

Overall this qualitative analysis indicates that there are both strengths and limitations in MentalRoBERTa's classification of suicide-related content. The model relies heavily on individual keywords, which suggests a lack of nuanced contextual understanding, where emotionally charged words override broader sentence meaning, resulting in bias and misclassification risks. Conversely, correctly classified instances show a more balanced distribution of importance scores, particularly in well-defined categories, where contextual clarity aids accuracy. These findings indicate that contextual ambiguity remains a challenge, limiting the model's real-world applicability in automated moderation and in detecting suicide-related content.

Interpretability: We conducted an N-gram extraction focusing on unigrams and bigrams. The results for bigrams are shown in Figure 3, while the unigram results are provided in Appendix D. First, we used NLTK to handle text preprocessing, such as removing stopwords and cleaning up the text by eliminating irrelevant terms and punctuation. To perform the N-gram extraction, we employed the *CountVectorizer* from Sklearn module. This tool allows us to tokenize the text and extract both unigrams (single words) and bigrams (pairs of consecutive words). By setting the *ngram_range* parameter to (1, 1) for unigrams and (2, 2) for bigrams, we instructed the *CountVectorizer* to focus on these specific N-gram types. After tokenization, we calculated the frequency of each unigram and bigram, then sorted them by their frequency to identify the most common terms and phrases in the dataset.

This visualization highlights the top bigrams across various suicidal content categories, each associated with distinct themes. The *Humorous use* category includes phrases like 'I'm gonna' and 'look like', often blending possible sarcasm with potentially distressing content. *Content not relevant* captures terms such as 'ready jump' and

[2]https://github.com/cdpierse/transformers-interpret

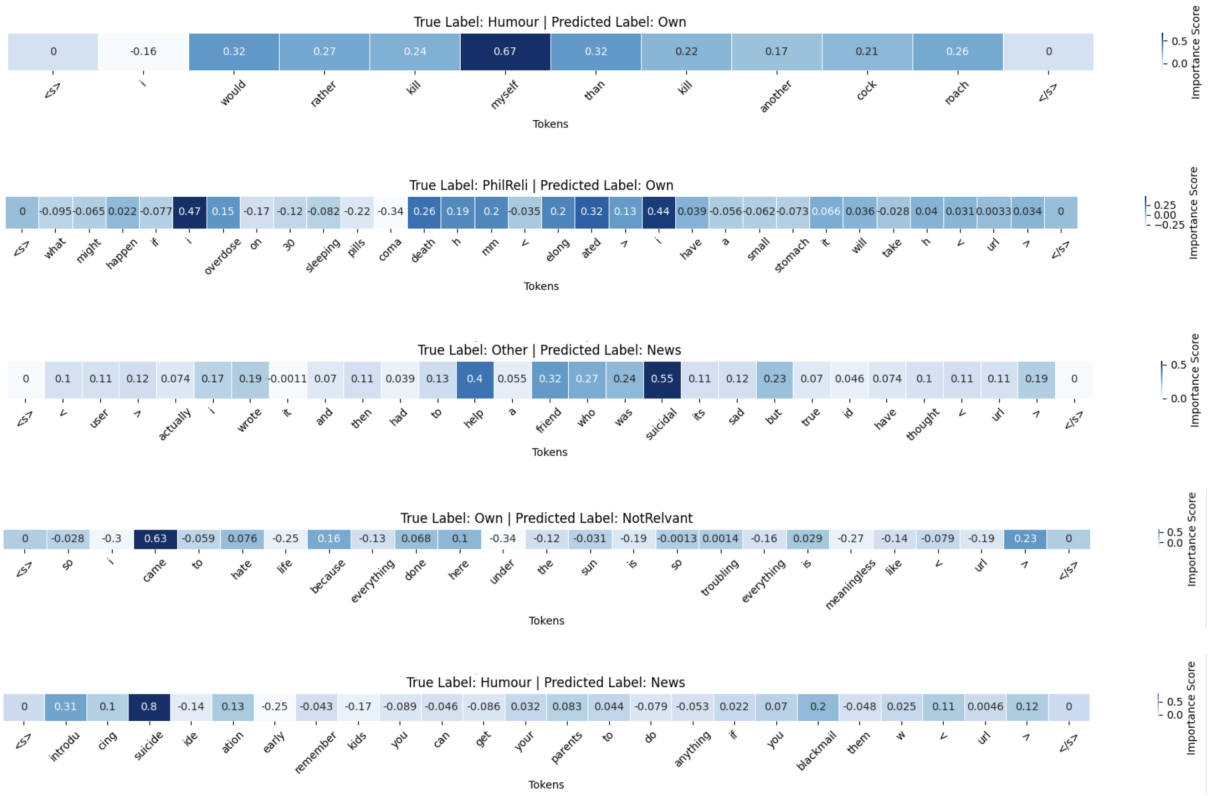

Figure 1: Token attribution scores for incorrectly classified tweets.

'life worth', which may reference life struggles without clear suicidal ideation. The *Expressing Own Suicidality* category prominently features self-referential terms like 'better dead' and 'tired living', indicating expressions of personal distress. The *Facts about suicide* category focuses on awareness-related terms such as 'mental health' and 'prevention day' reflecting informational content and social media campaigns. *Philosophical/Religious content* includes existential phrases like 'don't understand' and 'worth living', capturing discussions about the value of life. In the *News and Stories* category, bigrams such as 'mental commit' and 'thoughts URL' suggest references to suicide-related news articles or reports. The *Expressing worries about suicidality of others* category captures miscellaneous expressions, including empathetic phrases like 'sorry hear'. Lastly, the *Contact for suicide related help-seeking* category emphasizes help-seeking language, with phrases like 'need help' and 'talk need', along with references to hotlines and lifelines. Additionally, some phrases, such as 'I'm gonna' and 'don't wanna', appear across multiple categories, including *Humorous use*,*Expressing Own Suicidality*, and *Expressing worries about suicidality of others*, highlighting overlapping themes of distress and self-expression across different contexts.

Our qualitative analysis and findings underscore the limitations of traditional keyword-based approaches in suicide-related content detection and highlight the need for more sophisticated methods to better understand suicide-related language. This is particularly important, as the use of current approaches to detect risk can have significant and detrimental implications.

7 Conclusions and Future Directions

In this study, we expanded the existing TWISCO dataset, a multi-categorical suicide tweet dataset, and evaluated the performance of eight state-of-the-art language models, including both general-purpose and domain-specific models. Our findings highlight that domain-specific models outperform general-purpose counterparts as data size increases. Additionally, a content category-wise analysis showed improved performance in minority categories within the updated dataset. Furthermore, we conducted an interpretability and explainability analysis on the top-performing model, *MentalRoBERTa*, by examining token importance scores for both correctly classified and misclassified con-

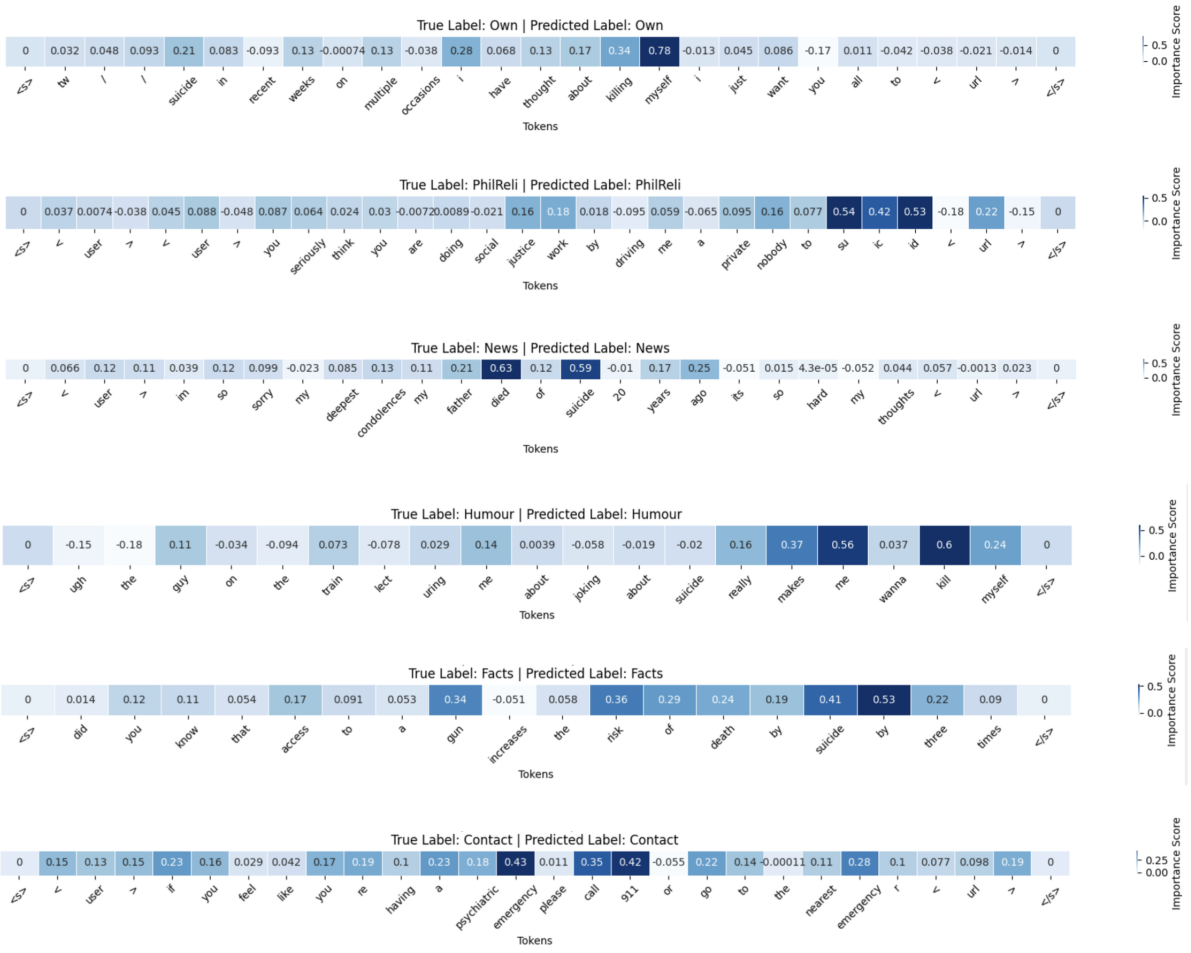

Figure 2: Token attribution scores for correctly classified tweets.

tent. Our analysis revealed that certain suicide-related terms receive high token importance scores, often overlooking crucial contextual nuances, such as humor or the personal expression of suicidal thoughts. We performed N-gram extraction per content category to identify top N-grams associated with each category. We found significant overlap across categories, particularly between *Humorous use* and *Expressing own Suicidality* posing challenges for context identification. However, categories such as *Contact for Suicide-Related Help-Seeking* and *News and Stories* exhibited distinct contextual patterns.

In the future, we aim to extend this work by leveraging Large Language Models (LLMs) with private training and comparing their performance against traditional language models. Additionally, we seek to explore whether incorporating additional context during training can further enhance model performance.

Limitations

Despite the advancements presented in this study, several limitations remain. First, dataset challenges persist, particularly the underrepresentation of certain categories, such as Contacts for Suicide-Related Help-Seeking and Expressing Worries About Others' Suicidality, which may limit the model's ability to generalize effectively across diverse suicidal content. Additionally, the downsampling of the *Content Not Relevant* category, while necessary for balance, may introduce biases that do not fully reflect real-world data distributions. Furthermore, model performance remains inconsistent, particularly in minority categories, with domain-specific models such as *MentalRoBERTa* demonstrating improvements primarily when sufficient training data is available. The study's interpretability analysis reveals an overreliance on emotionally charged keywords (e.g., 'kill', 'suicidal', 'hate'), often at the expense of contextual understanding, leading to misclassifications, especially in differ-

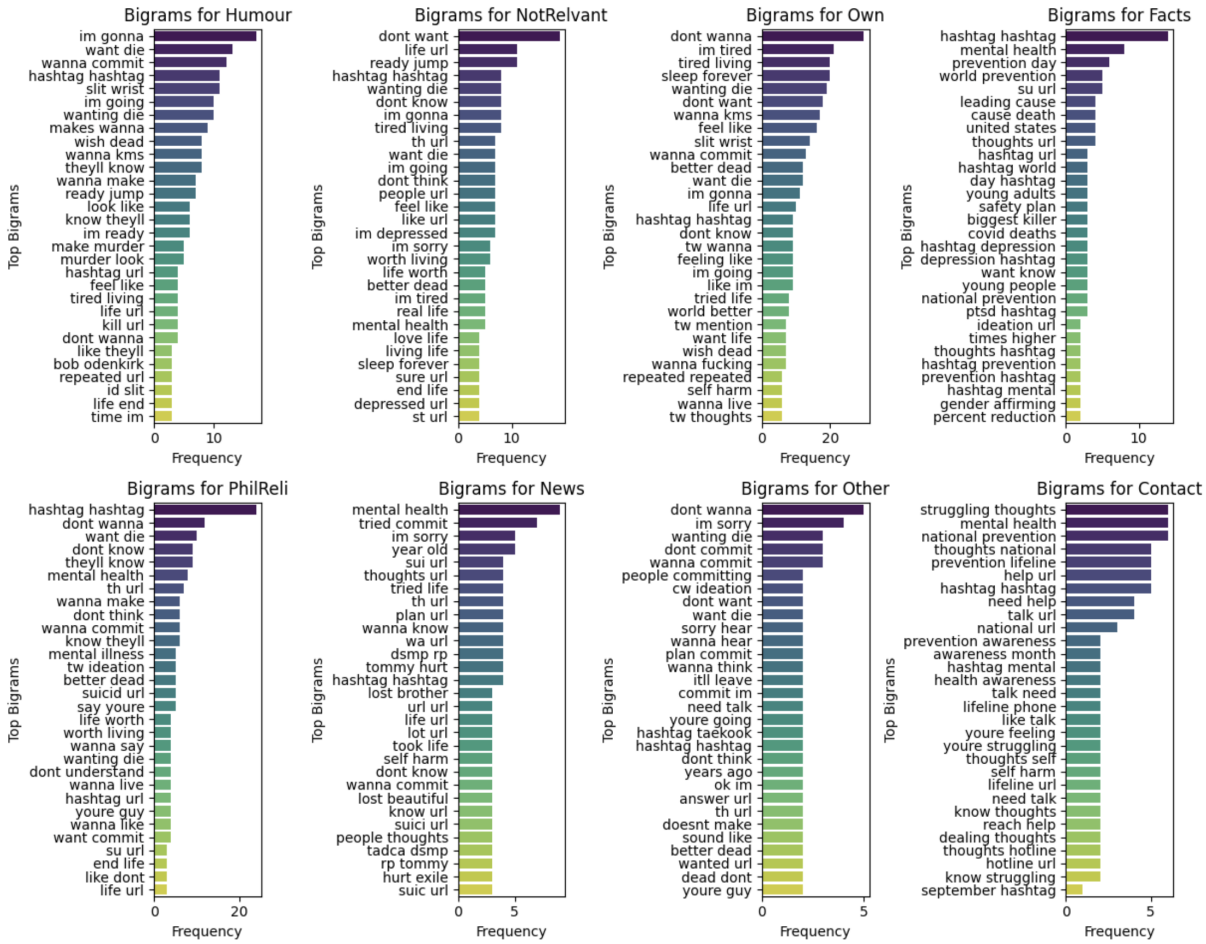

Figure 3: Bigrams extracted for each suicide content category

entiating humor, figurative language, and genuine suicidal intent. Lastly, the real-world applicability of the proposed methods remains constrained, as current models lack the necessary robustness for deployment in high-stakes scenarios. Additionally, while the study justifies the exclusion of Large Language Models (LLMs) due to ethical and computational concerns, this limits the ability to evaluate whether such models could provide superior contextual comprehension. Future research should explore context-aware modeling approaches, investigate privacy-preserving methods for training LLMs to enhance reliability and generalizability in suicide-related content detection tasks.

Ethics Statement

The study raises several ethical considerations regarding the use of LM's for detecting suicide-related content, particularly in terms of misclassification risks, bias, privacy, and real-world implications. Given the high-stakes nature of suicide related content detection, these concerns highlight the limitations of current approaches and the need for careful ethical considerations before deploying such models in practical applications:

- One of the most pressing ethical concerns in this study is the potential for misclassification, which can have serious consequences in real-world applications. LMs often struggle to differentiate between humor, figurative language, and genuine distress, leading to false positives and false negatives. Given these risks, this study underscores that LMs should not be used in isolation for crisis interventions and require mandatory human oversight.

- The study highlights the tendency of models to over-prioritize emotionally charged words such as 'kill', 'suicidal', and 'hate', often at the expense of contextual understanding. This raises concerns about bias in model predictions, particularly against certain demographic or linguistic groups that may use suicide-related language differently. Suicide-related

discussions vary across cultures, dialects, and social groups. A model trained primarily on Western, English-language social media posts may perform poorly when analyzing content from non-Western contexts or multilingual users, leading to disparities in detection accuracy (Schoene et al., 2025). Some groups, particularly those in LGBTQ+ communities, racial minorities, or neurodivergent individuals, may discuss mental health in unique ways that LMs fail to recognize. This could lead to over- or under-detection in these populations, affecting their access to mental health support and crisis intervention. This study finds that domain-specific models outperform general-purpose models only when sufficient training data is available, suggesting that minority categories may remain underrepresented. This could further exacerbate bias and misclassification risks, especially in real-world applications.

- The use of publicly available social media data to train suicide-related content detection models raises concerns related to privacy, consent, and data security. While social media provides valuable real-world data for training LMs, its use is particularly sensitive due to the personal and distressing nature of the content. Even when using publicly available data, there is a risk that user-generated content could be repurposed without individuals being aware, raising legal and ethical concerns regarding data ownership and exploitation.

- We acknowledge that current explainability techniques (e.g., attention visualization, feature attribution) provide only limited insight into how models classify suicide-related content. The lack of interpretability in LMS raises concerns regarding trust, accountability, and ultimately decision-making in real-world settings. For example, if LMs prioritize certain words over context, users and moderators may struggle to understand why certain posts are flagged. Furthermore, if explainability remains limited, human moderators or mental health professionals may find it difficult to effectively use and trust LM outputs.

Acknowledgements

References

Danai-Dionysia Athanasopoulou. 2024. Data protection in the era of generative artificial intelligence: Navigating gdpr compliance challenges in medical applications of chatgpt.

Salah Bouktif, Akib Mohi Ud Din Khanday, and Ali Ouni. 2025. Explainable predictive model for suicidal ideation during covid-19: Social media discourse study. *Journal of Medical Internet Research*, 27:e65434.

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*.

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.

Shangying Hua, Shuangci Jin, and Shengyi Jiang. 2024. The limitations and ethical considerations of chatgpt. *Data intelligence*, 6(1):201–239.

Md Rafiqul Islam, Md Kowsar Hossain Sakib, Shanjita Akter Prome, Xianzhi Wang, Anwaar Ulhaq, Cesar Sanin, and David Asirvatham. 2023. Machine learning with explainability for suicide ideation detection from social media data. In *2023 10th International Conference on Behavioural and Social Computing (BESC)*, pages 1–6. IEEE.

S Ji, T Zhang, L Ansari, J Fu, P Tiwari, and E Cambria. 2021. Mentalbert: publicly available pretrained language models for mental healthcare. arxiv. *Preprint posted online on December*, 29.

L Joseph and Levin Fleiss. 2023. *Statistical methods for rates and proportions*. Wiley-Blackwell.

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

Mikhail V Koroteev. 2021. Bert: a review of applications in natural language processing and understanding. *arXiv preprint arXiv:2103.11943*.

Jun Li, Xinhong Chen, Zehang Lin, Kaiqi Yang, Hong Va Leong, Nancy Xiaonan Yu, and Qing Li. 2022. Suicide risk level prediction and suicide trigger detection: A benchmark dataset. *HKIE Transactions Hong Kong Institution of Engineers*, 29(4):268–282.

Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. 2023. Trustworthy llms: A survey and guideline for evaluating large language models' alignment. *arXiv preprint arXiv:2308.05374*.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Seyedali Mohammadi, Edward Raff, Jinendra Malekar, Vedant Palit, Francis Ferraro, and Manas Gaur. 2024. Welldunn: On the robustness and explainability of language models and large language models in identifying wellness dimensions. *arXiv preprint arXiv:2406.12058*.

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*.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2023. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*.

Kamal Nigam, John Lafferty, and Andrew McCallum. 1999. Using maximum entropy for text classification. In *IJCAI-99 workshop on machine learning for information filtering*, volume 1, pages 61–67. Stockholm, Sweden.

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.

Adonias Caetano de Oliveira, Renato Freitas Bessa, and Ariel Soares Teles. 2024. Comparative analysis of bert-based and generative large language models for detecting suicidal ideation: a performance evaluation study. *Cadernos de Saúde Pública*, 40:e00028824.

Mahboobeh Parsapoor, Jacob W Koudys, and Anthony C Ruocco. 2023. Suicide risk detection using artificial intelligence: the promise of creating a benchmark dataset for research on the detection of suicide risk. *Frontiers in psychiatry*, 14:1186569.

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.

Bhanu Pratap Singh Rawat, Samuel Kovaly, Wilfred R Pigeon, and Hong Yu. 2022. Scan: Suicide attempt and ideation events dataset. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, volume 2022, page 1029. NIH Public Access.

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.

ABM Kamrul Islam Riad, Md Abdul Barek, Md Mostafizur Rahman, Mst Shapna Akter, Tahia Islam, Md Abdur Rahman, Md Raihan Mia, Hossain Shahriar, Fan Wu, and Sheikh Iqbal Ahamed. 2024. Enhancing hipaa compliance in ai-driven mhealth devices security and privacy. In *2024 IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC)*, pages 2430–2435. IEEE.

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 Schoene, Resmi Ramachandranpillai, Tomo Lazovich, and Ricardo Baeza-Yates. 2024. All models are wrong, but some are deadly: Inconsistencies in emotion detection in suicide-related tweets. In *Proceedings of the Third Workshop on NLP for Positive Impact*, pages 113–122.

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.

Annika Marie Schoene, John Ortega, Silvio Amir, and Kenneth Church. 2023. An example of (too much) hyper-parameter tuning in suicide ideation detection. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 1158–1162.

Annika Marie Schoene, John E Ortega, Rodolfo Joel Zevallos, and Laura Haaber Ihle. 2025. Lexicography

saves lives (lsl): Automatically translating suicide-related language. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3179–3192.

Vedant Vajre, Mitch Naylor, Uday Kamath, and Amarda Shehu. 2021. Psychbert: a mental health language model for social media mental health behavioral analysis. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1077–1082. IEEE.

Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. 2024. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, page 100211.

Liu Yinhan, Ott Myle, Goyal Naman, Du Jingfei, Joshi Mandar, Chen Danqi, Levy Omer, and Lewis Mike. 2019. Roberta: A robustly optimized bert pretraining approach (2019). *arXiv preprint arXiv:1907.11692*, pages 1–13.

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 Content label description

The following section provides more insights on each content label based on previous work by Schoene et al. (2022):

- Facts about suicidality: While factual information on suicide may appear neutral, research suggests it can be harmful to at-risk individuals seeking methods online. Websites discussing suicide facts often include how-to guides and even rank suicide methods based on lethality and associated pain .
- Philosophical or religious discussions of suicide: Discussions about suicide within philosophical or religious contexts may include judgmental perspectives or stigmatizing language, which can discourage individuals from seeking help.
- Contacts for suicide prevention and help-seeking: Some posts provide guidance on where to seek help and include links to support services and crisis resources.
- News reports, case studies, or stories: Narrative stories and reports by news outlets, activist groups, or individuals.
- Humorous references to suicide: The use of suicide-related phrases in a sarcastic or joking manner can lead to misclassification by suicide intent detection algorithms, potentially flagging content incorrectly.
- Content not relevant: Due to the nature of the dataset collection process, some content is unrelated to the task and should be classified accordingly.
- Expressing own suicidality: Individuals who openly express their own suicidal thoughts or feelings are often experiencing significant distress. Detecting such cases through an algorithmic approach and providing prevention resources could be beneficial.
- Expressing worries about the suicidality of other: Similar to those expressing their own suicidality, individuals who voice concerns about someone else's suicidal thoughts or behaviors may also be experiencing high levels of distress and worry.

B Figure: Category-wise performance analysis

C Training Details

The dataset is first divided into a training set (70%) and a test-validation set (30%), with the test and validation sets further split equally (50/50) while ensuring stratification on the target variable to maintain label distribution. The model is trained for 10 epochs, with a low learning rate of 1e-5 to prevent overfitting and ensure gradual optimization. Both training and evaluation are conducted with a batch size of 8. To mitigate overfitting, a weight decay of 0.01 is applied. Evaluation is performed at the end of each epoch, and model checkpoints are saved periodically to capture the best-performing model based on validation results, with the best model automatically loaded at the end of training. We used *Discovery cluster* for all our computing experiments.

D Unigrams for Suicide Content Categories

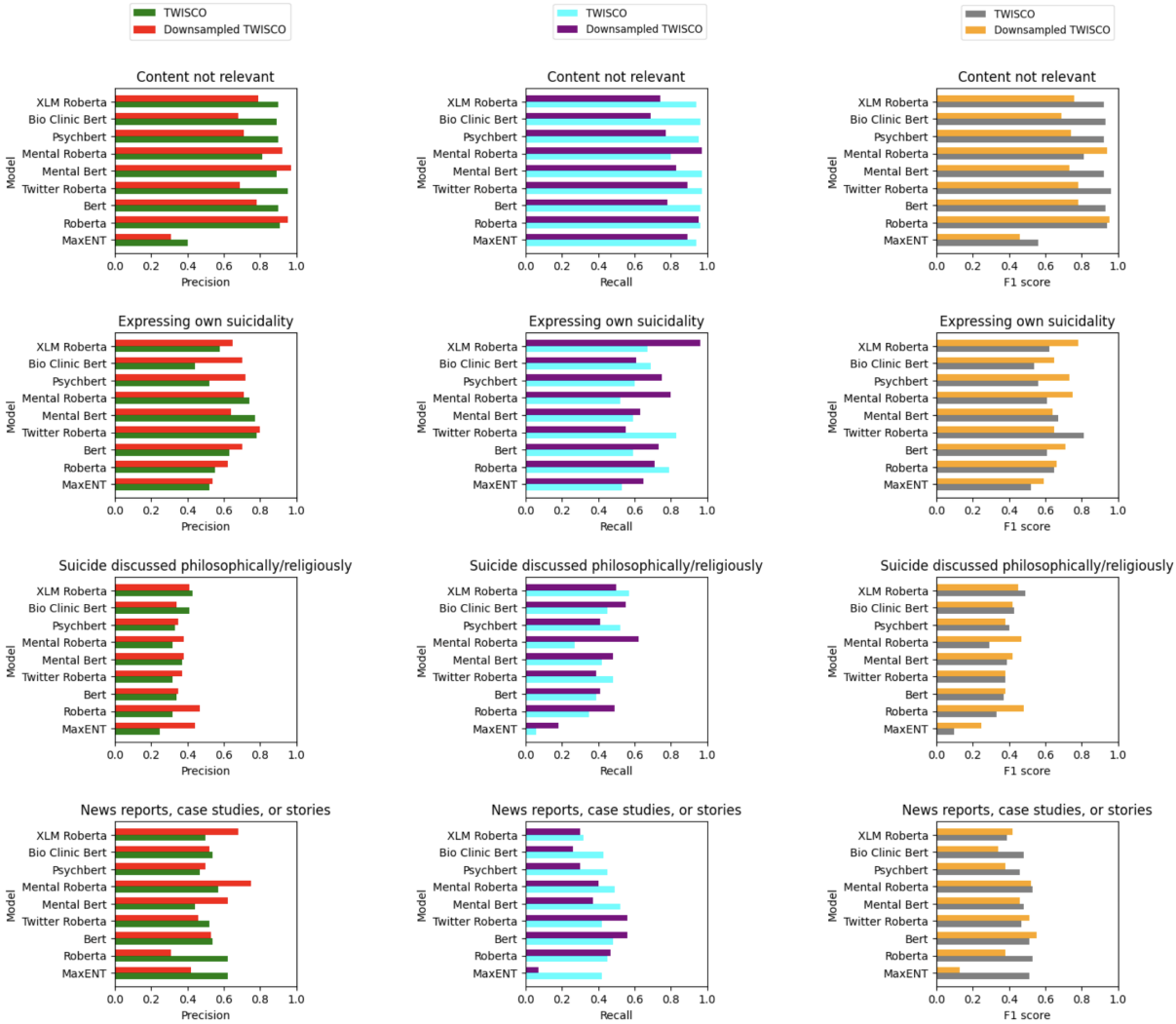

Figure 4: Precision, recall, and F1 scores for categories: *content not relevant*, *expressing own suicidality*, *suicide discussed philosophically/religiously*, and *news reports, case studies or stories*

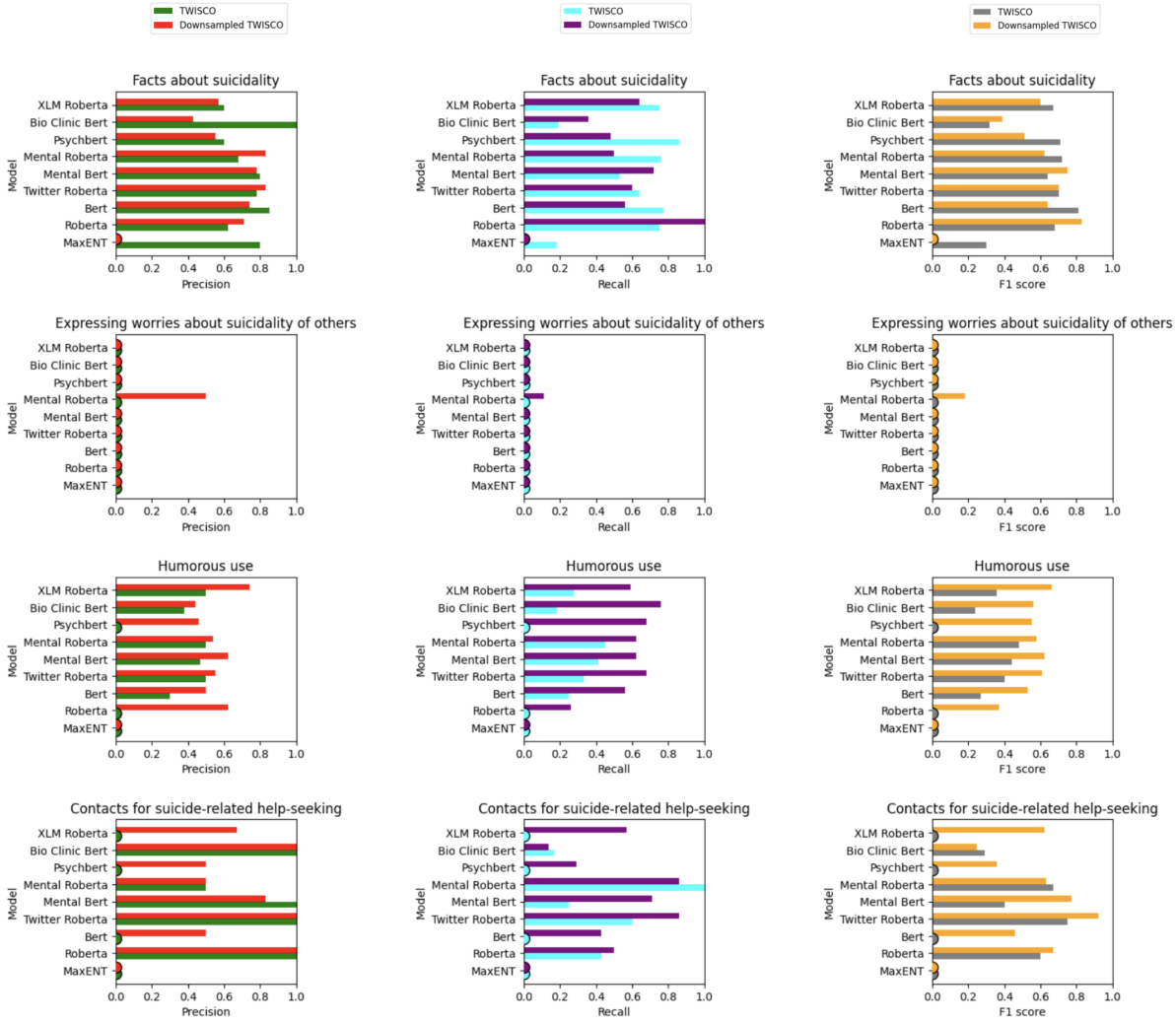

Figure 5: Precision, recall, and F1 scores for suicide-related content categories: *facts about suicidality*, *expressing worries about suicidality of others*, *humorous use*, and *contacts for suicide-related help seeking*

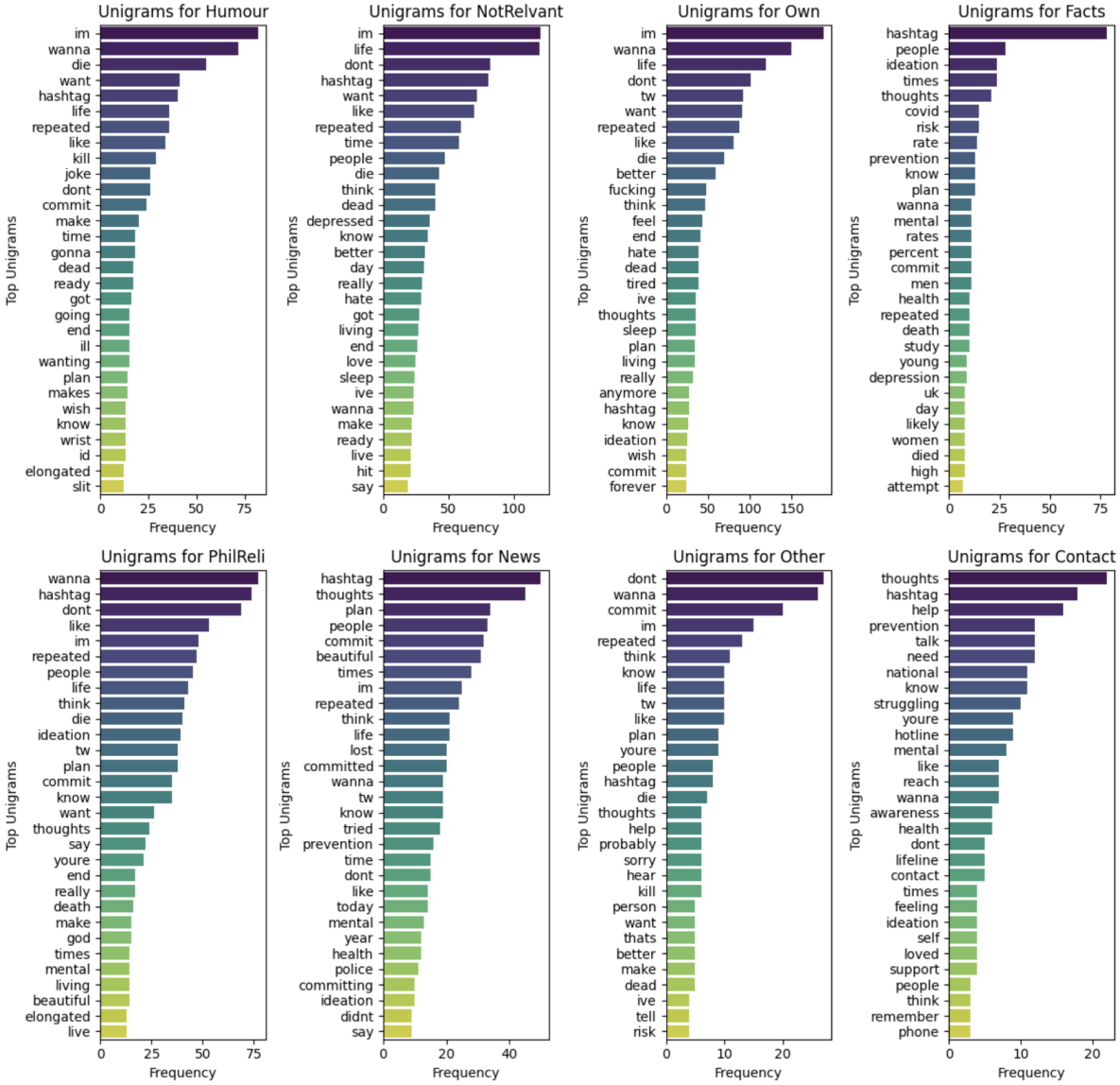

Figure 6: unigrams extracted for each suicide content category