TONY: an open-source *TO*olkit for *N*lp in ps*Y*chology

Federico Ravenda[★], **Sofia Irene Ravenda**[♡],
Volodymyr Karpenko[★], **Daniele Montagnani**[★, △],
Andrea Raballo[★, ♣, *], **Antonietta Mira**[★, ◇, *]
{name.lastname}@usi.ch, s.ravenda@campus.unimib.it
[★] Euler Institute, Università della Svizzera italiana, [♡] Università degli Studi Milano-Bicocca,
[△] University of Pavia, [♣] Cantonal Socio-Psychiatric Organization , [◇] University of Insubria

Abstract

The growing demand for Mental Health (**MH**) services highlights the need for scalable computational tools, yet progress in computational psychology is hindered by scarce sensitive data, complex assessment procedures, and high technical barriers. While language is a well-established marker of different MH conditions, existing NLP solutions are often fragmented, closed-source, or difficult to use, limiting their adoption in interdisciplinary research. We present **TONY**, an open-source, python **TO**olkit for **N**LP in clinical ps**Y**chology. TONY bridges traditional psycholinguistic analysis and modern NLP by combining interpretable lexical features with state-of-the-art lightweight transformer models within a unified and easy-to-use framework. This hybrid approach enables robust and transparent text analysis without relying on large-scale models or closed-source software. TONY is designed for researchers and practitioners working at the intersection of NLP and MH, facilitating collaboration across disciplines. Compared to the few existing systems, TONY offers a more comprehensive and exhaustive solution, reducing the barrier to entry through a unified, modular, and reproducible pipeline that integrates classical and neural approaches in a single open framework. The toolkit is released under an open-source license and is evaluated through multiple MH-related datasets, demonstrating its flexibility and effectiveness in low-resource settings[1,2].

1 Introduction

Human language is a complex and multifaceted medium, laden with subtle signals that go far beyond literal meaning. A single sentence can simultaneously convey emotion, social dynamics, cognitive states, and implicit attitudes, information that is invaluable for understanding human behavior at scale. A growing body of work in computational social science, human-computer interaction, and social computing has sought to surface these signals automatically: modeling empathy expression in therapeutic conversations (Hu et al., 2025; De Grandi et al., 2025), detecting hate speech and subtle forms of online harassment (Plaza-del Arco et al., 2023; Nozza et al., 2019), tracking longitudinal changes in emotional well-being through personal narratives (Chakraborty et al., 2025) or inferring personality traits from writing style (Gjurković et al., 2021).

These signals become crucial in the mental health (hereafter **MH**) domain. The escalating global demand for MH services has underscored a critical need for innovative solutions that can scale beyond traditional clinical boundaries (Bajaj, 2024) and language stands at the center of this challenge. Language has long been recognized as both a fundamental human signature and a significant manifestation of MH conditions: the words people choose, the syntactic structures they favor, and the emotional tone they project are all windows into their psychological state (He et al., 2026). The intricate relationship between linguistic patterns and psychological states has captured the attention of clinicians and linguists for decades, establishing a rich foundation for interdisciplinary collaboration. Yet despite this shared interest, computational linguistics and clinical psychology have often evolved as parallel worlds, each governed by distinct methodological frameworks and priorities. This disconnect is particularly acute in the realm of computational psychology, where the development of AI-driven diagnostic tools is constrained by the scarcity of sensitive clinical data and the complexity of psychological assessment. Bridging this gap, i.e., building tools that are simultaneously rigorous enough for clinical insight and scalable enough for large-scale research, remains an open challenge.

[*] Shared last authorship
[1] Repository Github for TONY package:
[2] TONY documentation website https://fede-stack.github.io/tony.github.io/

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), pages 660–671

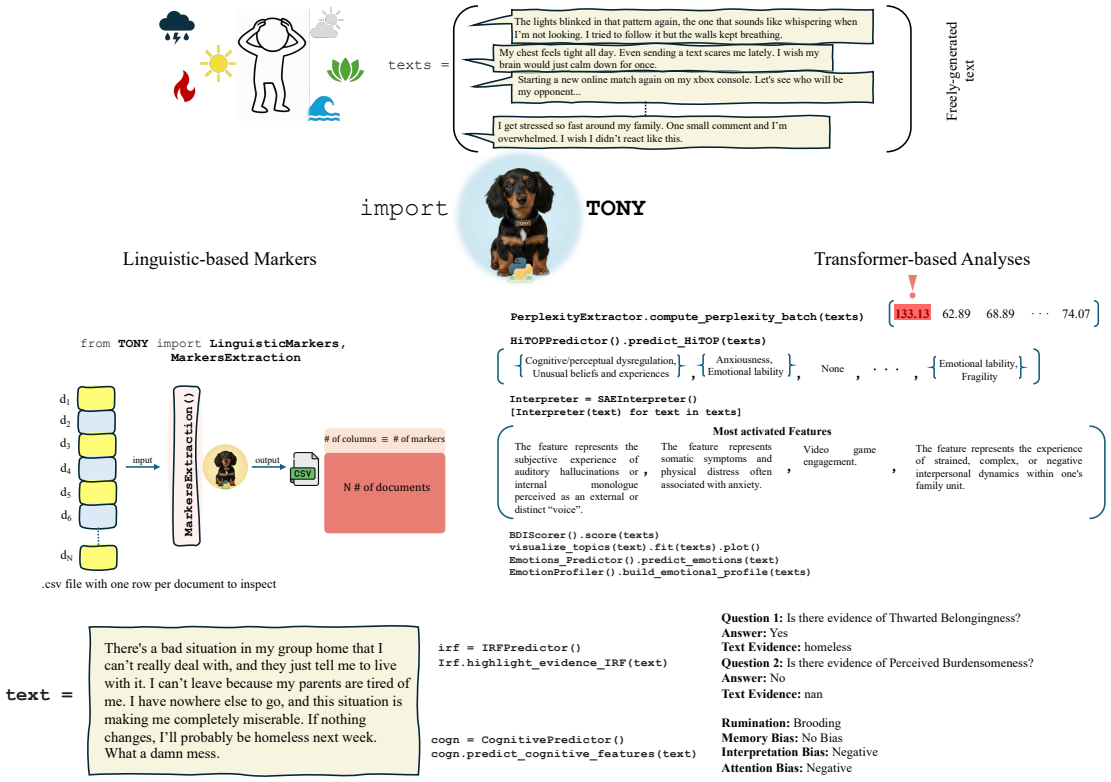

Figure 1: Overview of TONY's two main components. *Left*: the lexicon-based module processes freely-generated texts and outputs a structured `.csv` file of linguistic markers (one row per document). *Right*: the transformer-based module provides perplexity estimation, HiTOP trait prediction, interpretable latent feature activation via a Sparse Autoencoder, BDI-II completion from user post histories , topic modeling, emotion prediction, and detection of IRF and cognitive features with textual evidence.

Lexicon-based approaches have long served as a backbone for this kind of analysis. Tools like LIWC (Linguistic Inquiry and Word Count) (Tausczik and Pennebaker, 2010) enable researchers to quantify psychological and emotional constructs in text by counting words belonging to predefined categories, such as sadness, health, or positive emotion. LIWC and similar lexicons-based tools offer undeniable advantages: they are computationally lightweight, highly interpretable, and validated across a wide range of domains. Yet lexicon-based methods carry significant limitations. First, they operate purely at the word level, and are therefore blind to contextual meaning, negation, irony, and the compositional semantics that shape how language conveys mental states. Second, widely used resources like LIWC are closed-source, limiting reproducibility and extensibility. Third, these lexicons tend to be narrow in scope and frequently lack coverage of domains of emerging research interest, forcing researchers to curate and validate new word lists on an ad-hoc basis. The rapid advancement of NLP and the rise of transformer-based language models offer a compelling path beyond these constraints.

To address these gaps, we present TONY (*TO*olkit for *N*lp in ps*Y*chology), a Python package designed to support the linguistic analysis of text in the context of MH research. TONY is built around two complementary pillars: **(1)** a comprehensive suite of *linguistic-based features*, spanning categories such as lexical diversity and sophistication, syntactic complexity, pronoun and verb tense distribution, negation frequency, emotion and sentiment scores, affective dimensions (valence, arousal, dominance), discourse cohesion, and cognitive and social process markers, which allow for fast, interpretable, and reproducible analyses; and **(2)** a set of *transformer-based modules* that leverage state-of-the-art pre-trained and low-resource models, specifically adapted to the psychological domain, to capture contextual, semantic, and psychologically grounded signals beyond the reach of word-level methods, enabling the extraction of clinically meaningful features that reflect not only surface linguistic patterns but also deeper cognitive and emotional states. By combining the transparency of lexical approaches with the expressive power of modern NLP, TONY aims to lower the barrier to rigorous, large-scale linguistic analysis in

MH research while remaining fully open-source and easily extensible.

2 Background

Different tools exist for extracting linguistic features from text, yet none has been designed specifically for the demands of MH research. Here we briefly survey the most commonly used approaches to position our contribution.

Lexicon-based methods represent the established tradition. LIWC (Tausczik and Pennebaker, 2010) is the most widely adopted tool, offering 62 syntactic, topical, and emotional categories validated across a broad range of domains. The General Inquirer (Stone et al., 1966) extends topical coverage but with fewer emotional categories, while tools such as EmoLex (Yavuz, 2020), ANEW (Bradley and Lang, 1999), and SentiWordNet (Baccianella et al., 2010) focus on richer emotional and sentiment inventories. Empath (Fast et al., 2016) builds on this tradition by providing a larger set of categories and the ability to generate new ones on demand via unsupervised language modeling, combining the interpretability of word lists with the flexibility of data-driven methods. While effective for large-scale analyses, these approaches share a fundamental limitation: they operate at the word level, remaining blind to context, negation, and the compositional semantics that characterize MH discourse. Other tools like SpeechGraphs (Mota et al., 2012) take structural approaches, extracting graph-based metrics from speech transcripts, but remain focused on specific phenomena and disorder types.

The advent of transformer architectures has fostered a growing body of work applying both general-purpose and domain-adapted language models to MH contexts (Guo et al., 2024), enabling the capture of rich contextual and psychologically grounded representations that go well beyond what lexicon-based tools can encode. A representative example is the work of Fernández-Iglesias et al. (2024), where the authors released DepressMind, an automated system that analyzes social media posts to estimate and monitor linguistic signals associated with depression, benchmarking them against the established dimension of Beck Depression Inventory II (BDI-II, a psychometric tool for depression screening). However, such tools remain scattered across individual studies, each confined to a specific task or condition, and lack a unified, accessible interface for researchers. TONY addresses this fragmentation by providing a single, extensible, and fully open-source framework that integrates *linguistic* features and *transformer-based* analyses, purpose-built for the linguistic study of MH data.

3 TONY - The First Comprehensive TOol for NLP in PsYchology

TONY spans a broad range of modules for computing lexical features, neural-based analyses, and visualization. Each of these modules is described in the following sections.

3.1 Linguistic Feature Module

This module extracts a broad set of linguistically motivated features from raw text, organized into nine categories. **(i) Lexical features** capture vocabulary richness and sophistication, including Type-Token Ratio (TTR) as a measure of lexical diversity, word repetition rate as a proxy for rumination, average word frequency derived from large-scale corpus statistics, and word prevalence based on function word usage. **(ii) Syntactic features** quantify structural complexity through subordination and coordination rates, average sentence length, mean dependency distance computed over spaCy dependency trees (Honnibal et al., 2020), incomplete sentence ratio, and a composite sentence complexity score. **(iii) Morphosyntactic and POS features** include the distribution of major part-of-speech categories, nouns, verbs, adjectives, adverbs, prepositions, auxiliary verbs, conjunctions, and interjections, as well as morphological markers such as indicative and subjunctive mood ratios and singular versus plural noun usage, and dependency-level features capturing subject and direct object rates per sentence. **(iv) Stylistic features** track pronoun usage across first, second, and third person, a well-established marker of depression and self-focus (Zimmermann et al., 2017), alongside verb tense distribution, past-to-future tense ratio, negation frequency, and question and exclamation ratios. **(v) Emotion and sentiment features** leverage the NRC Emotion Lexicon (Mohammad and Turney, 2013) to produce normalized scores across eight basic emotions (joy, sadness, anger, fear, disgust, surprise, anticipation, and trust), complemented by sentiment polarity and a VADER-based (Hutto and Gilbert, 2014) compound intensity score. **(vi) Affective features** approximate valence, arousal, and dominance dimensions by mapping NRC emotion

scores onto the VAD affective space. **(vii) Domain-specific lexicon features** compute the normalized frequency of clinically motivated word categories, including absolutist language, death-related vocabulary, anxiety, sadness, and anger words, somatic and body-related terms, and achievement or grandiosity language, each grounded in the psycholinguistic literature on specific MH conditions. **(viii) Cohesion features** measure inter-sentence connectivity through lexical overlap, a Jaccard-based cohesion score, and discourse connective usage - patterns particularly relevant for detecting thought disorder and disorganized speech. Finally, **(ix) cognitive, social, and referential features** quantify the relative frequency of insight, causation, certainty, and tentative language, social reference words computed over lemmatized tokens, and Named Entity Recognition (NER) based rates of person and temporal references. A complete list of all extracted features is provided in Table 3 in the Appendix.

3.2 Transformer-based Modules

These modules provide a suite of analyses that leverage pre-trained and fine-tuned language models specifically adapted to the MH domain. In the following, we illustrate the main functionalities of the modules through concrete code snippets and their corresponding outputs, using the following example text:

```
text = 'Some days I keep living, even though I feel completely alone in the world'
```

(i) HiTOP Traits Prediction. `TONY` predicts psychopathological trait dimensions from the Hierarchical Taxonomy of Psychopathology (HiTOP) (Kotov et al., 2022) directly from text. The HiTOP framework organizes psychopathology into a hierarchy of transdiagnostic traits, such as anxiousness, emotional lability, or cognitive/perceptual dysregulation, that represent the core dimensions underlying MH conditions, making their automatic detection from language a particularly informative source of clinically grounded features. The underlying model is a fine-tuned LLM trained on PersonalityDBench (Ravenda et al., 2026), a dataset of Reddit posts and comments spanning a broad range of MH-related subreddits, where each post is annotated with one or more HiTOP traits, or with none when no trait is identifiable in the content. The choice of a generative model for this multi-label classification task was motivated by the poor performance observed when fine-tuning encoder-based models, which struggled to capture the nuanced and overlapping nature of HiTOP trait assignments. Fine-tuned model variants are publicly available at huggingface.co/FritzStack/models under the prefix `HiTOP`.

```
hitop = HiTOPPredictor()
hitop.predict_HiTOP(text)
# Output
# Anhedonia, Withdrawal,
    Depressivity
```

(ii) Interpersonal Risk Factors. `TONY` detects the two core interpersonal risk factors of the Interpersonal Theory of Suicide (Van Orden et al., 2010), Thwarted Belongingness (TBE) and Perceived Burdensomeness (PBU). TBE refers to the painful feeling of being disconnected from others and lacking a sense of belonging, while PBU captures the perception of being a liability or burden to those around oneself. According to the Interpersonal Theory of Suicide, the co-occurrence of these two factors is a critical precursor to suicidal ideation, making their automatic detection from text a clinically significant task for early risk identification. The module use an LLM fine-tuned on the Interpersonal Risk Factors of Mental Disturbance dataset (Garg et al., 2023) and not only predicts the presence of each factor but also highlights the supporting textual evidence, providing interpretable outputs for both clinical and research use. The default model is a 4-bit quantized Qwen3 8B, with additional fine-tuned variants available at huggingface.co/FritzStack/models under the prefix `IRF`.

```
predictor = IRFPredictor()
predictor.highlight_evidence_IRF(
↪text)
# Output
# Question 1: Is there evidence of
    Thwarted Belongingness?
# Answer: Yes
# Text Evidence: feel completely
    alone
# Question 2: Is there evidence of
    Perceived Burdensomeness?
# Answer:  No
# Text Evidence: nan
```

(iii) Cognitive Features Prediction. TONY incorporates a module for detecting cognitive distortion patterns directly from text, drawing on the theoretical framework introduced by Agarwal et al. (2025), where authors established that four cognitive dimensions - *rumination, memory bias, interpretation bias, and attention bias* - are clinically meaningful markers that significantly differentiate

depression relapse from non-relapse users in social media data, validating long-standing cognitive theories of depression in a computational setting. Each dimension captures a distinct form of cognitive distortion: rumination refers to repetitive and passive focus on negative thoughts; memory bias reflects the tendency to selectively recall negative past experiences; interpretation bias characterizes the inclination to assign negative meanings to ambiguous situations; and attention bias captures the preferential orientation toward negative or threatening stimuli. TONY implements the automatic prediction of these four dimensions through the 'CognitivePredictor' class, which uses fine-tuned LLMs on the ReDepress dataset (models are available at huggingface.co/FritzStack/models under the prefix `COGN`) to produce structured, per-text predictions:

```
cogn = TONY.CognitivePredictor()
cogn.predict_cognitive_features(
↪text)
# Rumination: Brooding
# Memory Bias: No Bias
# Interpretation Bias: Negative
# Attention Bias: Negative
```

The output provides an interpretable cognitive profile for each text, capturing maladaptive thinking styles that are otherwise difficult to operationalize through lexical methods alone, making the module particularly suited for longitudinal analyses of early MH risk signals.

All the default models for HiTOP traits, IRF, and cognitive features predictions *are designed to be accessible without high-end hardware*: they run efficiently on a T4 GPU, freely available through Google Colab, making them suitable for researchers without dedicated computational resources. Additionally, MLX-converted versions with Q4 quantization are provided for all models, enabling smooth execution on Apple Silicon chips (M-series) for users working in local environments.

(iv) SAE Interpretation applies a Sparse Autoencoder (SAE) trained on more than 50,000 Reddit posts spanning from casual conversation to MH-focused communities. The SAE operates on text representations produced by `qwen3-embedding-4b` and learns to decompose each dense embedding into a sparse set of interpretable latent features following O'Neill et al. (2024); full implementation details are provided in Appendix A. Three configurations of increasing capacity are available in `TONY` (SAE16, SAE32, SAE64), with SAE64 set as the default. Each feature is automatically labelled via a two-stage LLM pipeline (see Appendix A). Given an input text, `TONY` identifies the most strongly activated features and returns their natural language descriptions, as shown in Figure 1 and in the following example:

```
interpreter = SAEInterpreter()
interpreter(text, top_k=1, # show
    only the most activated feature
↪ sae='SAE64')
# {'text': 'Some days I keep living
    , even though I feel completely
     alone in the world', 'sae': '
    SAE64', 'features': [{'
    feature_id': 11104, 'label': '
    The feature represents the
    subjective experience of
    chronic, pervasive, or
    existential loneliness', 'sae':
     'SAE64', 'rank': 1}]}
```

(v) Emotion Prediction (`predict_emotions`, `build_emotional_profile`) predicts fine-grained emotion labels for individual texts using a transformer-based classifier, and constructs longitudinal emotional profiles from sequences of user posts, enabling the tracking of affective states over time. An illustrative example of an emotional profile is provided in Figure B in the Appendix.

(vi) Cluster Visualization `TONY` extracts and visualizes latent topics from a collection of texts following a BERTopic-inspired pipeline (Grootendorst, 2022), enabling exploratory analysis of thematic patterns across large corpora. The key difference lies in the use of UMAP* as discussed in Di Noia et al. (2026), in which the number of components and neighbors used to project observations into a lower-dimensional space is determined adaptively by the framework, rather than being fixed a priori, and clustering is subsequently performed in that reduced space via KMeans. Optionally, topic representations and labels can be further refined by a lightweight large language model running locally, making the step suitable for resource-constrained or offline environments without sacrificing interpretability. Figure 2 illustrates an example applied to the DAIS-C corpus (Delgaram-Nejad et al., 2023), a small, specialised collection of spoken language in which speakers diagnosed with schizophrenia and those with no self-reported psychiatric or neuroleptic history were interviewed on the same topics. Only the interviewees' responses are considered, and topics along with their representations are computed by fixing an arbitrary number of 10 topic clusters.

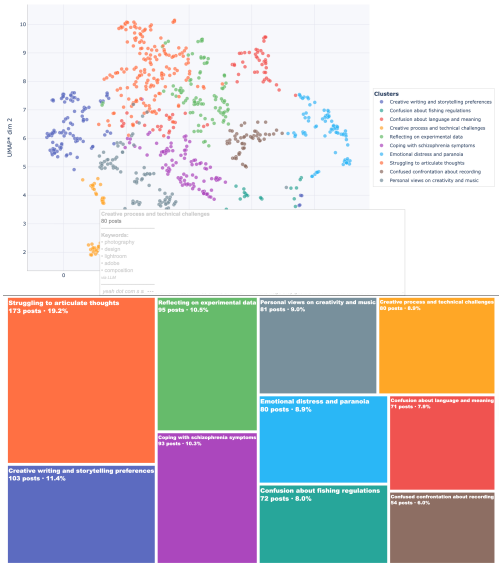

Figure 2: Topic visualization on the DAIS-C corpus. *(Top)* Interactive scatter plot of the UMAP* two-dimensional embedding, with each point representing an interview response colored by cluster assignment. *(Bottom)* Mosaic plot displaying LLM-refined topic labels alongside the percentage of texts assigned to each of the 10 inferred clusters.

(vii) Embeddings and BDI-II Completion produces document-level semantic representations using SBERT-based models and provides an adaptive Retrieval-Augmented Generation (aRAG) (Ravenda et al., 2025) pipeline for automatically completing the BDI-II questionnaire from a user's post history. For the retrieval component, we adapted `all-mpnet-base-v2` to the BDI-II domain using the DepreSym dataset (Pérez et al., 2025) (see Section 4 for a description of the dataset), fine-tuning it with triplet loss to optimize the retrieval of posts relevant to each BDI-II item and choice. The core of the pipeline is an adaptive retrieval mechanism that, for each BDI-II item j and choice l, retrieves a variable-size set of relevant posts from the user's history $\aleph$. Formally, the Retrieved Relevant Posts are defined as: $\mathbf{RRP}_{ijl} = \{p_i\}_{p_i \in \aleph(iq_{jl})}$, where $\aleph(iq_{jl})$ represents the adaptive neighborhood of size k^* for item j and choice l. Unlike standard RAG approaches that fix the number of retrieved documents k a priori, the adaptive neighborhood k^* is determined dynamically based on the semantic density of the user's post history relative to each item, allowing the pipeline to retrieve more evidence when available and less when the signal is sparse. The retrieved posts are then passed to a generative LLM to produce a structured BDI-II response. The pipeline is compatible with both OpenAI-compatible clients and HuggingFace-hosted models, allowing flexible deployment across different computational environments.

4 Datasets

To evaluate the performance of TONY's different functionalities, several datasets were employed. In particular, the quality of the linguistic features extracted by `TONY` was assessed using three classification datasets in the MH domain. The first is **DepreSym** (Pérez et al., 2025), a collection of 21,580 Reddit posts annotated for relevance to BDI-II symptoms. The second is the **Interpersonal Risk Factors (IRF)** dataset (Garg et al., 2023), a collection of social media posts (train/test 1972/1057 annotated to assess whether each post contains evidence of Thwarted Belongingness (TBE) and Perceived Burdensomeness (PBU), along with the corresponding text spans highlighted as supporting evidence. The third is the **Depression Severity** dataset (Naseem et al., 2022), where each post (3,553 in total) is annotated with a depression intensity level drawn from the following categories: minimum, mild, moderate, and severe. The IRF dataset was also used to fine-tune lightweight LLMs on the task of detecting the presence of the two interpersonal risk factors and identifying the text spans in which they appear. Finally, the **eRisk 2021** dataset (Parapar et al., 2021), developed for the Measuring the Severity of Depression Signs task, was employed. Each user (80 in total) is paired with their complete Reddit post history and a self-reported BDI-II questionnaire. The objective is to predict the user's BDI-II responses as accurately as possible given their post history.

5 Experiments

Lexical Features Evaluation. The quality of the lexical features was evaluated by using them as input to MH classification tasks on three datasets and comparing the results with features generated by LIWC-22 (a commercial tool) and Empath (open source). For each tool, lexical embeddings were extracted from the text as encoded by the respective package and fed into a classifier. Following the methodology adopted in MTEB Benchmark (Muennighoff et al., 2023), we used Logistic Regression as classifier. Since the Depression Severity dataset involves ordinal categorical labels (minimum, mild, moderate, severe), we employed an ordinal vari-

method	ACC	F1	P	R
DepreSym (Pérez et al., 2025)				
LIWC	0.770	0.580	**0.7157**	0.494
Empath	0.759	0.283	0.345	0.245
TONY	**0.783**	**0.582**	0.609	**0.560**
IRF (TBE/PBU) (Garg et al., 2023)				
LIWC	(0.671, **0.780**)	(0.703, **0.619**)	(0.704, **0.670**)	(0.701, **0.575**)
Empath	(0.621, 0.709)	(0.662, 0.451)	(0.653, 0.538)	(0.672, 0.388)
TONY	(**0.696**, 0.7759	(**0.721**, 0.588)	(**0.735**, 0.687)	(**0.706**, 0.514)
MentalBERT	(0.751, 0.783)	(0.767, 0.628)	(**0.780**, 0.642)	(0.774, 0.658)
gpt-5.4-mini	(0.768, 0.661)	(0.813, 0.527)	(0.736, 0.462)	(0.910, 0.612)
qwen-72b	(0.760, 0.689)	(0.809, 0.551)	(0.722, 0.495)	(**0.920**, 0.551)
`TONY.IRFPredictor()`				
	(**0.781**, **0.880**)	(**0.823**, **0.810**)	(0.745, **0.789**)	(**0.920**, **0.831**)
Depression Severity (Naseem et al., 2022)				
LIWC	0.733	**0.371**	**0.538**	**0.350**
Empath	0.727	0.212	0.207	0.250
TONY	**0.736**	0.332	0.500	0.321

Table 1: Classification results across TONY, LIWC-22, and Empath on three MH datasets. For IRF, values are reported as (TBE, PBU) pairs. `TONY.IRFPredictor()` is compared against transformers-based baselines. Bold values indicate the best result per metric.

ant of Logistic Regression (Harrell Jr, 2015) to better respect the natural ordering of the target classes, while standard Logistic Regression was used for the remaining datasets. For the DepreSym and Depression Severity datasets, 10-fold cross-validation was applied, while for IRF we adopted the train/test split provided by the original authors. Results show that `TONY` and LIWC are the two best-performing tools, while Empath lags behind. Notably, on DepreSym, `TONY` outperforms LIWC on all metrics except precision; on IRF for the TBE task, `TONY` systematically outperforms both competitors; and on Depression Severity, `TONY` achieves the best accuracy overall. These results suggest that `TONY`'s features are highly effective at capturing the psychological dimensions of text, and are often more discriminative than those of LIWC - a commercial tool - for MH prediction. Additionally, for the IRF dataset, the training split was used to fine-tune different LLMs (available at huggingface.co/FritzStack/models under the prefix `IRF`) to produce the output shown in Figure 1. As reported in Table 1, the quantized 4-bit Qwen-8B model obtained through this process substantially outperforms both the lexical feature baseline and MentalBERT (Ji et al., 2022) - a BERT variant specifically adapted to the MH domain and fine-tuned on the same training set - as well as `gpt-5.4-mini` and `Qwen-2.5-72B-instruct` used as zero-shot classifiers (the prompt used for zero-shot classification is reported in the Appendix C).

Completing the BDI-II from Reddit Histories. As anticipated, DepressMind (Fernández-Iglesias et al., 2024) recently introduced a framework for automatically completing the BDI-II questionnaire from a user's Reddit post history, retrieving relevant posts via cosine similarity and applying an NLI entailment model to verify whether each post supports a given BDI-II response. While computationally lightweight, this method yields limited performance on the eRisk benchmark. More recently, Ravenda et al. (2025) proposed an aRAG-based approach that adaptively retrieves the most relevant posts for each item and performs zero-shot prediction via a generative LLM, achieving stronger results. TONY implements this through the `BDIScorer` class, which takes as input a list of posts for a given user and returns a 21-dimensional vector of predicted BDI-II item scores. As shown in Table 2, all TONY configurations substantially outperform both DepressMind and the best eRisk 2021 participants across all of the metrics.

method	DHCR	ADODL	AHR	ACR
eRisk 2021				
DepressMind	0.3375	0.7383	0.2601	0.6319
participants	0.2196	0.7586	0.3107	0.6555
`TONY.BDIScorer()`				
`gemma 27B`	**0.563**	**0.859**	**0.373**	**0.734**
`lfm-2.2-6B`	0.538	0.841	0.344	0.704
`gemma 4B`	0.450	0.825	0.334	0.701

Table 2: Results on the eRisk 2021 Task: Measuring the Severity of Depression Signs. **Bold** values represent the best performance on a specific metric.

6 Conclusions and Future Work

`TONY` is the first open-source toolkit specifically designed for the linguistic analysis of MH text, uniquely integrating interpretable linguistic-based features with state-of-the-art transformer-based analyses within a single unified framework. Unlike existing tools, often limited to lexical approaches, restricted to specific conditions, or closed-source, `TONY` is fully open-source and purpose-built for the MH domain, lowering the barrier to entry for researchers working at the intersection of NLP and clinical psychology. As the field evolves, `TONY` will be actively maintained and updated to incorporate new models, fine-tuned resources, and additional feature modules, ensuring that the toolkit remains aligned with the latest advances in NLP and computational psychology.

7 Ethical Considerations

TONY is intended solely for research purposes and must not be used as a substitute for clinical diagnosis or professional mental health assessment. All datasets used in this work were derived from publicly available Reddit data; no personally identifiable information was collected or stored, and all data were handled in accordance with standard anonymization practices in computational social science research. The automatic detection of suicide risk indicators, interpersonal risk factors, and psychopathological traits is inherently sensitive, and outputs should be interpreted only by qualified professionals within appropriate clinical or research contexts in critical contexts. We acknowledge that both lexical and transformer-based models may reflect biases present in their training data, and we encourage users to validate TONY's outputs on their specific target populations before drawing conclusions. The BDI-II completion pipeline is designed exclusively as a research tool to support large-scale studies and should not be used for individual-level clinical decision-making.

8 Limitations & Broader Impact

TONY has the potential to advance research at the intersection of NLP and clinical psychology by providing a unified, open-source framework that lowers technical barriers for researchers. By making state-of-the-art tools for mental health text analysis freely accessible, TONY can facilitate large-scale studies previously constrained by the cost of commercial tools such as LIWC or by the fragmentation of existing resources. In the longer term, tools like TONY may contribute to the earlier detection of mental health risk signals, supporting the development of scalable, language-based screening systems that complement traditional clinical workflows. At the same time, deploying automated mental health analysis tools outside controlled research settings entails significant risks, including potential misuse for surveillance, stigmatization of vulnerable populations, and overreliance on model outputs in high-stakes decisions. We therefore strongly encourage the research community to engage with such tools responsibly, in close collaboration with mental health professionals and with careful attention to the ethical implications of each specific use case.

Some limitations should be acknowledged. The lexicon-based module relies on predefined word lists that may not generalize equally well across writing styles, demographics, or clinical populations. The transformer-based models are trained predominantly on English Reddit data, which may limit their applicability to other languages, platforms, or clinical contexts. Finally, all outputs should be interpreted as research signals rather than clinical assessments, and their validity should be verified on the specific target population before drawing conclusions.

Acknowledgements

This research has partly been founded by the SNSF (Swiss National Science Foundation) grant 200557.

References

Aakash Kumar Agarwal, Saprativa Bhattacharjee, Mauli Rastogi, Jemima S. Jacob, Biplab Banerjee, Rashmi Gupta, and Pushpak Bhattacharyya. 2025. ReDepress: A cognitive framework for detecting depression relapse from social media. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 34664–34682, Suzhou, China. Association for Computational Linguistics.

Stefano Baccianella, Andrea Esuli, Fabrizio Sebastiani, and 1 others. 2010. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Lrec*, volume 10, pages 2200–2204. Valletta.

Simar Bajaj. 2024. How students and grandparents could solve the global mental-health crisis. *Nature*, 635(8039):540–542.

Margaret M. Bradley and Peter J. Lang. 1999. Affective norms for english words (anew): Instruction manual and affective ratings.

Suchandra Chakraborty, Sudeshna Jana, Manjira Sinha, and Tirthankar Dasgupta. 2025. Self-state evidence extraction and well-being prediction from social media timelines. In *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, pages 279–286.

Alessandro De Grandi, Federico Ravenda, Andrea Raballo, and Fabio Crestani. 2025. The emotional spectrum of llms: Leveraging empathy and emotion-based markers for mental health support. In *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, pages 26–43.

Oliver Delgaram-Nejad, Dawn Archer, Gerasimos Chatzidamianos, Louise Robinson, and Alex Bartha. 2023. The dais-c: A small, specialised, spoken, schizophrenia corpus. *Applied Corpus Linguistics*, 3(3):100069.

Antonio Di Noia, Federico Ravenda, and Antonietta Mira. 2026. A general framework for adaptive non-parametric dimensionality reduction. *Scientific Reports*.

Ethan Fast, Binbin Chen, and Michael S Bernstein. 2016. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 4647–4657.

Roque Fernández-Iglesias, Marcos Fernández-Pichel, Mario Ezra Aragón, and David E Losada. 2024. Depressmind: A depression surveillance system for social media analysis. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 35–43.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Muskan Garg, Amirmohammad Shahbandegan, Amrit Chadha, and Vijay Mago. 2023. An annotated dataset for explainable interpersonal risk factors of mental disturbance in social media posts. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11960–11969, Toronto, Canada. Association for Computational Linguistics.

Matej Gjurković, Vanja M Karan, Iva Vukojević, Mihaela Bošnjak, and Jan Šnajder. 2021. Pandora talks: Personality and demographics on reddit. In *Proceedings of the ninth international workshop on natural language processing for social media*, pages 138–152.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Zhijun Guo, Alvina Lai, Johan H Thygesen, Joseph Farrington, Thomas Keen, and Kezhi Li. 2024. Large language models for mental health applications: systematic review. *JMIR mental health*, 11(1):e57400.

Frank E Harrell Jr. 2015. Ordinal logistic regression. In *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*, pages 311–325. Springer.

Jiaman He, Marta Micheli, Damiano Spina, Dana McKay, Johanne R Trippas, and Noriko Kando. 2026. Characterizing personality from eye-tracking: The role of gaze and its absence in interactive search environments. *arXiv preprint arXiv:2601.08287*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, Adriane Boyd, and 1 others. 2020. spacy: Industrial-strength natural language processing in python.

He Hu, Yucheng Zhou, Juzheng Si, Qianning Wang, Hengheng Zhang, Fuji Ren, Fei Ma, Laizhong Cui, and Qi Tian. 2025. Beyond empathy: Integrating diagnostic and therapeutic reasoning with large language models for mental health counseling. *arXiv preprint arXiv:2505.15715*.

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. Mentalbert: Publicly available pretrained language models for mental healthcare. In *proceedings of the thirteenth language resources and evaluation conference*, pages 7184–7190.

Roman Kotov, David C Cicero, Christopher C Conway, Colin G DeYoung, Alexandre Dombrovski, Nicholas R Eaton, Michael B First, Miriam K Forbes, Steven E Hyman, Katherine G Jonas, and 1 others. 2022. The hierarchical taxonomy of psychopathology (hitop) in psychiatric practice and research. *Psychological medicine*, 52(9):1666–1678.

Saif M Mohammad and Peter D Turney. 2013. Nrc emotion lexicon. *National Research Council, Canada*, 2:234.

Natalia B Mota, Nivaldo AP Vasconcelos, Nathalia Lemos, Ana C Pieretti, Osame Kinouchi, Guillermo A Cecchi, Mauro Copelli, and Sidarta Ribeiro. 2012. Speech graphs provide a quantitative measure of thought disorder in psychosis. *PloS one*, 7(4):e34928.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. Mteb: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037.

Usman Naseem, Adam G Dunn, Jinman Kim, and Matloob Khushi. 2022. Early identification of depression severity levels on reddit using ordinal classification. In *Proceedings of the ACM web conference 2022*, pages 2563–2572.

Debora Nozza, Claudia Volpetti, and Elisabetta Fersini. 2019. Unintended bias in misogyny detection. In *Ieee/wic/acm international conference on web intelligence*, pages 149–155.

Charles O'Neill, Christine Ye, Kartheik Iyer, and John F. Wu. 2024. Disentangling dense embeddings with sparse autoencoders. *Preprint*, arXiv:2408.00657.

Javier Parapar, Patricia Martín-Rodilla, David E Losada, and Fabio Crestani. 2021. erisk 2021: pathological gambling, self-harm and depression challenges. In *European Conference on Information Retrieval*, pages 650–656. Springer.

Anxo Pérez, Marcos Fernández-Pichel, Javier Parapar, and David E Losada. 2025. Depresym: A depression

symptom annotated corpus and the role of large language models as assessors of psychological markers: A. pérez et al. *Language Resources and Evaluation*, 59(3):2737–2762.

Flor Miriam Plaza-del Arco, Debora Nozza, and Dirk Hovy. 2023. Respectful or toxic? using zero-shot learning with language models to detect hate speech. In *The 7th workshop on online abuse and harms (woah)*, pages 60–68.

Federico Ravenda, Seyed Ali Bahrainian, Daniele Montagnani, Antonietta Mira, and Andrea Raballo. 2026. Personalitydbench: A dataset for personality disorders – from modeling to controlled generation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*. Accepted to ACL 2026 Main Conference.

Federico Ravenda, Seyed Ali Bahrainian, Andrea Raballo, Antonietta Mira, and Noriko Kando. 2025. Are llms effective psychological assessors? leveraging adaptive rag for interpretable mental health screening through psychometric practice. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8975–8991.

Philip J Stone, Dexter C Dunphy, and Marshall S Smith. 1966. The general inquirer: A computer approach to content analysis.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

Kimberly A Van Orden, Tracy K Witte, Kelly C Cukrowicz, Scott R Braithwaite, Edward A Selby, and Thomas E Joiner Jr. 2010. The interpersonal theory of suicide. *Psychological review*, 117(2):575.

Mehmet Can Yavuz. 2020. Analyses of character emotions in dramatic works by using emolex unigrams. In *Proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it 2020)*, pages 472–477.

Johannes Zimmermann, Timo Brockmeyer, Matthias Hunn, Henning Schauenburg, and Markus Wolf. 2017. First-person pronoun use in spoken language as a predictor of future depressive symptoms: Preliminary evidence from a clinical sample of depressed patients. *Clinical psychology & psychotherapy*, 24(2):384–391.

A SAE Implementation Details

Embeddings from `qwen3-embedding-4b` ($d = 2560$) are normalised to zero mean and unit variance before being passed to the SAE. The model decomposes each embedding $\mathbf{x} \in R^d$ into a sparse latent representation $\mathbf{z} \in R^n$, where $n > d$ is the dictionary size, via a top-k activation function that retains only the k largest activations and sets the rest to zero, avoiding the shrinkage bias of ℓ_1 penalisation. The decoder reconstructs the input as $\hat{\mathbf{x}} = W_d\mathbf{z} + \mathbf{b}$, and the training objective combines a normalised reconstruction loss with an auxiliary loss on dead latents (ghost gradients, $\alpha = 1/32$) to prevent latent collapse:

$$\mathcal{L} = \frac{1}{d}\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 + \alpha\,\mathcal{L}_{aux}$$

Three configurations are trained for 13,200 steps (lr $= 10^{-4}$), where n denotes the number of latent dimensions: SAE16 (k=16, n=2d), SAE32 (k=32, n=4d), and SAE64 (k=64, n=6d). After training, features with activation frequency below 5×10^{-3} are discarded. The remaining features are interpreted via a two-stage pipeline using `google/gemini-flash-3.1-lite`: an *Interpreter* receives the ten most-activating texts alongside non-activating samples and generates a natural language label; a *Predictor* assigns a confidence score in $[-1, +1]$ to a new held-out set. Features with Pearson $r < 0.3$ between predicted confidence and ground-truth activation are excluded.

B Emotions as Markers of specific Mental Health Condition

De Grandi et al. (2025) demonstrate that distinct emotional distributions are systematically associated with specific MH conditions, establishing emotional profiles as meaningful linguistic markers. `TONY` implements this analysis through the `predict_emotions` and `build_emotional_profile` functions. To illustrate this, we computed emotional profiles on the Depression Severity dataset (Naseem et al., 2022), comparing posts labeled as minimum and severe depression. As shown in Figure 3, the two groups exhibit markedly different emotional distributions: posts with severe depression show substantially higher levels of *ashamed*, *anxious*, and *afraid*, while minimum-severity posts are characterized by higher *grateful*, *excited*, and *joyful* scores. The lower panel reports the results of `TONY.test_hypothesis()`, which performs statistical group comparisons across emotional dimensions, confirming, as expected, that emotions such as *excited*, *proud*, *grateful*, *lonely*, *joyful*, and *ashamed* differ significantly between depressed individuals and controls (p<0.05).

C Prompts for zero-shot IRF classification

Prompt:

```
You are a clinical NLP expert specializing in suicide risk assessment.
Classify social media posts for two interpersonal risk factors from Joiner's Interpersonal-Psychological Theory of Suicide (IPTS):
THWARTED BELONGINGNESS (TBE) — label 1 if the post expresses:
- Loneliness, social isolation, or feeling disconnected from others
- Feeling ignored, invisible, rejected, or excluded by people around them
- Absence of meaningful relationships, feeling unloved or unwanted
- Longing for connection, feeling no one cares or understands
- Estrangement from family, friends, or community
Label 0 if the post shows social connection, support, or no mention of belonging/isolation.
PERCEIVED BURDENSOMENESS (PBU) — label 1 if the post expresses:
- Believing oneself to be a burden, liability, or source of pain to others
- Feeling that others (family, friends, society) would be better off if the person were gone or dead
- Guilt over consuming resources, needing help, or causing problems for others
- A sense of being worthless, useless, or harmful to those around them
- Self-blame for the suffering of others
Label 0 if there is no expressed belief of being a burden or harmful to others.
Important: A post can be 1 for both, either, or neither. Base your judgment only on what is explicitly or strongly implicitly stated.
Respond ONLY with valid JSON, no other text.
```

D Complete List of Lexicon-Level Features

Table 3 provides a complete overview of all features extracted by `TONY`'s lexicon-level module, organized by category.

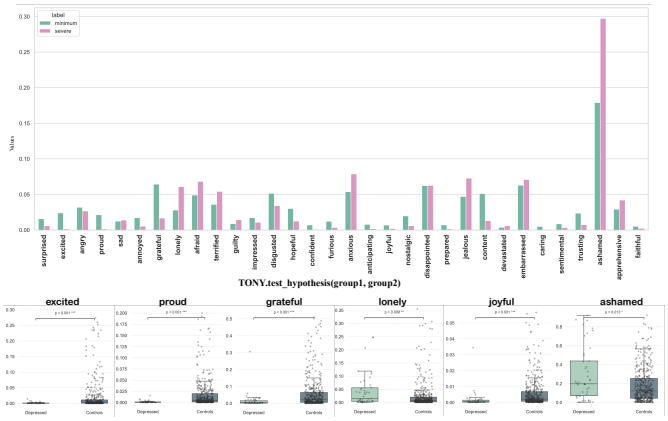

Figure 3: Top: emotional profile comparison between minimum and severe depression posts from the Depression Severity dataset (Naseem et al., 2022). Bottom: statistical group comparisons (`TONY.test_hypothesis()`) between depressed individuals and controls across six discriminative emotions, with significance levels indicated :$^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.001$.

Category	**Feature**	**Description**
Lexical	`lexical_diversity`	Type-Token Ratio (TTR)
	`lexical_sophistication`	Mean/std word frequency (wordfreq)
	`word_prevalence`	Function word ratio
	`repetition_rate`	Inverse TTR, proxy for rumination
Syntactic	`sentence_complexity`	Weighted sub/coordination score
	`subordination_rate`	Subordinate clause frequency
	`coordination_rate`	Coordinate clause frequency
	`mean_dependency_distance`	Mean dependency tree distance (spaCy)
	`incomplete_sentence_ratio`	Proportion of sentences < 3 words
	`average_sentence_length`	Mean words per sentence
Stylistic	`pronoun_usage`	First, second, and third person rates
	`verb_tense_distribution`	Past, present, and future tense ratios
	`past_future_ratio`	Log-ratio of past vs future tense usage
	`negation_frequency`	Negation word rate
	`question_ratio` / `exclamation_ratio`	Sentence-ending punctuation ratios
Emotion & Sentiment	`emotion_scores`	8 NRC emotions (normalized)
	`sentiment_polarity`	NRC positive/negative ratio
	`sentiment_intensity`	VADER compound score
	`affect_scores`	Valence, arousal, dominance (VAD space)
Cohesion	`cohesion_score`	Jaccard similarity across adjacent sentences
	`lexical_overlap`	Word overlap across adjacent sentences
	`connectives_usage`	Discourse connective rate
Cognitive & Social	`cognitive_processes`	Insight, causation, certainty, tentative rates
	`social_processes`	Social reference word count
	`readability_index`	Flesch Reading Ease (Flesch, 1948) (normalized)
	`graph_connectedness` / `semantic_coherence`	Word co-occurrence graph metrics
Domain Lexicons	`absolutist_word_frequency`	Absolutist language rate
	`death_word_frequency`	Death-related word rate
	`anxiety_word_frequency`	Anxiety lexicon rate
	`sadness_word_frequency`	Sadness lexicon rate
	`anger_word_frequency`	Anger lexicon rate
	`body_word_frequency`	Somatic/body lexicon rate
	`achievement_word_frequency`	Grandiosity lexicon rate
Morphosyntactic & NER	`pos_frequencies`	Preposition, auxiliary, adverb, conjunction rates
	`extended_pos`	Noun, verb, adjective, interjection rates
	`morphological_features`	Indicative/subjunctive mood, singular/plural ratios
	`dependency_features`	Subject and object rates per sentence
	`ner_features`	Person and temporal entity reference rates

Table 3: Complete list of features extracted by TONY's lexicon-level module, organized by category.