QCU The Star: Journal of Science, Engineering, and Information Technology
January-June 2026 | Vol. 4, No. 1
https://doi.org/10.64807/8122fq52

Detecting Suicidal Thoughts in Social Media: A Proposed NLP and Chatbot Framework Informed by a Preliminary Survey

Dhennise Puso[1] and Ma. Jackielyn Orbino
[1]Quezon City University
Quezon City, Philippines

Abstract. This conceptual paper presents a proposed natural language processing (NLP) and chatbot framework for recognizing potentially concerning language in social media posts and directing users toward human support. A preliminary survey provides context on respondents' platform use, encounters with posts expressing depressed thoughts, responses to such posts, and awareness of support resources. Demographic and platform-use summaries cover seven respondents, while three response items contain five tabulated answers each. Four of seven respondents were aged 25-34; all seven reported using Facebook, and five reported using Instagram. Of the five tabulated answers to the resource-awareness question, four indicated no awareness of support resources or helplines. The proposed workflow links authorized text collection, preprocessing, contextual classification, and a conversational interface that offers support information and routes users toward appropriate services. The survey informs initial design priorities, particularly accessible resource information and a clear route to human help. The contribution is a conceptual framework and a descriptive account of the preliminary survey; subsequent implementation research should examine classification performance, response quality, usability, privacy, and referral outcomes.

Keywords: natural language processing; social media; suicidal ideation; chatbot; conceptual framework; preliminary survey

Received: *October 21, 2025*
Revised: *April 13, 2026*
Accepted: *May 4, 2026*
Published: *June 30, 2026*

Introduction

Social media provides spaces where people discuss difficult experiences and express distress. A concerning post may be seen by someone who wants to help but does not know how to respond or where to find support information. This creates a design problem that involves both language processing and access to human assistance.

Natural language processing offers methods for examining patterns in text. Arowosegbe and Oyelade (2023) reviewed applications of NLP to detecting and preventing suicidal ideation and identified opportunities for computational approaches in this area. However, the meaning of a post depends on its context. Negative sentiment, references to suicide, and a statement of current personal intent are different forms of communication and should not be treated as interchangeable.

A conversational interface could help present support information after potentially concerning language is identified. Balcombe (2023) emphasizes the importance of human involvement, ethical responsibility, and attention to bias when considering chatbots in digital mental health. For the present framework, the intended role of the chatbot is to offer information and facilitate access to human support, rather than to diagnose a person or determine treatment.

Related work and conceptual basis

Prior work addresses distinct parts of the problem. De Choudhury et al. (2016) studied shifts from mental-health discussion to expressions of suicidal ideation in Reddit communities. Shing et al. (2018) examined expert, crowdsourced, and machine assessments of risk in online posts. These studies show why the choice of corpus, assessment unit, and annotation procedure matters; their findings cannot simply be transferred to another platform or to a different language community.

Conversational support involves a separate question. Fitzpatrick et al. (2017) evaluated a specific conversational agent in a randomized trial involving 70 young adults over two weeks. That intervention study differs from classifying social-media posts and does not establish the effectiveness of the framework proposed here. Together with the reviews by Arowosegbe and Oyelade (2023) and Balcombe (2023), these studies motivate a modular design in which language assessment, resource presentation, and human referral have separate responsibilities and evaluation criteria.

The contribution of this paper is a preliminary organization of those functions, informed by a small awareness survey. It does not claim a new classification algorithm or a demonstrated improvement over existing systems. The unresolved design question is how context-sensitive text assessment can be connected to voluntary, appropriate support without converting a computational signal into an unsupported judgment about a person.

Purpose and scope

This paper describes a proposed workflow that connects NLP-based text assessment with a chatbot interface. It also summarizes a preliminary survey of social-media use, encounters with posts expressing depressed thoughts, and awareness of support resources. The survey provides context for design priorities. The technical workflow is presented as a proposal, and the empirical results reported here are descriptive survey findings.

Approach

Conceptual development

The proposed framework organizes the functions discussed in the manuscript into four stages: authorized collection of social-media text, preprocessing and representation, contextual identification of potentially concerning content, and presentation of support information through a chatbot. The proposed architecture distinguishes the text-processing task from the human-support task so that a computational flag does not automatically become a clinical judgment.

Preliminary survey

The preliminary survey asked about respondents' age, social-media platform use, encounters with posts expressing depressed thoughts, actions taken in response, and awareness of support resources or helplines. Age and platform-use tables contain seven respondents. The three tables on encounters, responses, and resource awareness each contain five tabulated answers. Percentages in this paper are calculated from the counts for each question, using seven or five as appropriate.

Platform-use responses include more than one platform per respondent; consequently, those percentages are not expected to total 100%. The survey summaries are presented as counts and percentages. They describe the respondents' reported experiences and are not used to infer population prevalence or relationships between platform use and mental-health awareness.

Analytical boundaries

The age distribution and the platform-use counts provide context for the proposed design. They do not identify which age group has the highest risk of depression or suicide. Likewise, an answer about encountering a post expressing depressed thoughts is not evidence that the respondent, or the person who wrote the post, has a diagnosed condition. The distinctions between reported experience, observed language, and clinical assessment guide the interpretation of the findings.

Preliminary survey findings

Table 1

Age distribution of respondents

Age group	n	%
18-24	3	42.9
25-34	4	57.1
Total	7	100.0

Note. N = 7. All tabulated respondents fall within the two age groups shown.

Four respondents were aged 25-34 and three were aged 18-24. This describes the composition of the small survey group. It does not establish the distribution of mental-health concerns in these age groups.

Table 2

Reported social-media platform use

Platform	n	%
Facebook	7	100.0
Instagram	5	71.4
TikTok	3	42.9
X / Twitter	2	28.6
LinkedIn	2	28.6

Note. N = 7. Respondents could report more than one platform; percentages therefore exceed 100% when added.

All seven respondents reported using Facebook, while five reported using Instagram. The proposed use of text from X should therefore be treated as a technical design choice rather than as a preference established by this survey. A later implementation study would need to consider the platform actually used, its available interfaces, and the population it aims to serve.

Table 3

Encounters with posts expressing depressed thoughts

Response	n	%
Yes	3	60.0
No	2	40.0
Maybe	0	0.0
Total	5	100.0

Note. The question was: “Have you ever come across posts expressing depressed thoughts on social media?” Percentages use the five tabulated answers.

Three of the five tabulated answers indicated an encounter with posts expressing depressed thoughts. Two indicated no such encounter. These responses concern exposure to posts; they do not show whether respondents can identify suicidal intent or whether the authors of the posts received assistance.

Table 4

Reported actions in response to concerning posts

Action	n	%
Reached out privately	3	60.0
Reported the post to the platform	0	0.0
Encouraged professional help	0	0.0
Encouraged contact with friends or family	2	40.0
Other	0	0.0
Total	5	100.0

Note. Percentages use the five tabulated answers to the response question.

Private contact was the most frequently tabulated action, followed by encouragement to speak with friends or family. The counts indicate an inclination toward interpersonal support in this group. They do not establish the quality, appropriateness, or outcome of the assistance provided.

Table 5

Awareness of support resources or helplines

Response	n	%
Yes	1	20.0
No	4	80.0
Maybe	0	0.0
Total	5	100.0

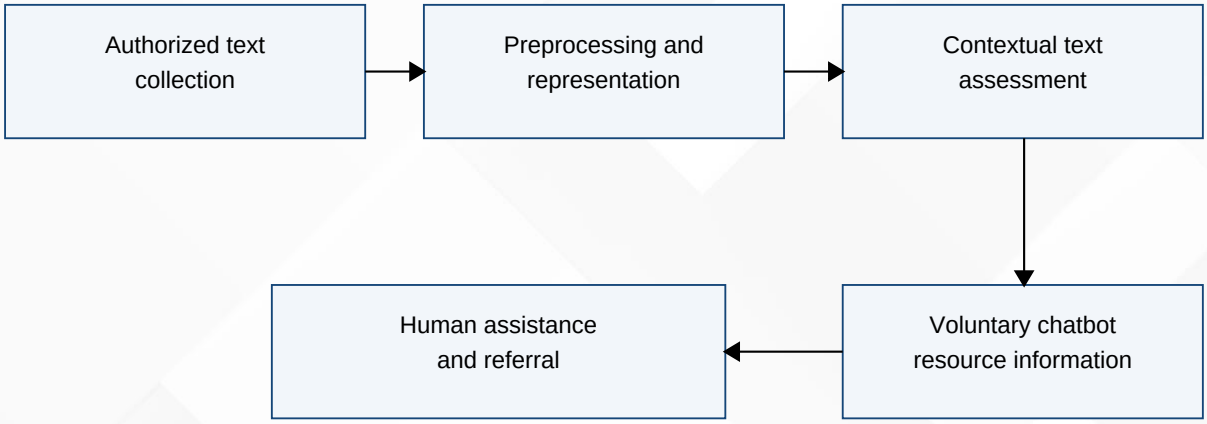
Note. Percentages use the five tabulated answers. The item asks about awareness of resources or helplines that support people dealing with mental-health issues.

Four of five tabulated answers indicated no awareness of support resources or helplines. Within this preliminary group, the finding supports giving resource information a visible place in the proposed interface. A resource directory would need to be relevant to the user's location and maintained by responsible personnel. The survey does not establish whether a chatbot is the most effective way to provide that information.

Proposed NLP and chatbot framework

Figure 1

Proposed workflow connecting text assessment with support information



Note. This diagram represents the proposed architecture. A text flag is a prompt for contextual review; access to support is voluntary. Human oversight, privacy controls, and error monitoring apply across all stages.

Authorized text collection

The collection stage would obtain social-media text through an authorized platform interface or another permitted research source. The original design identifies X as a possible source. Access permissions, the purpose of collection, retention periods, and the treatment of identifiable information would need to be defined before collection begins. The survey respondents and a future social-media text corpus are separate data sources.

Preprocessing and text representation

Preprocessing would prepare text for analysis through normalization, tokenization, and, where appropriate, lemmatization. Links, repost markers, and special characters would be handled according to a documented rule. Cleaning should preserve language features that affect meaning, including negation and relevant emoji or punctuation. The same preprocessing rules would need to be applied to training and evaluation text.

The representation stage could combine lexical features, word or sentence representations, and contextual information. Terms associated with distress would provide cues for further assessment rather than a sufficient basis for classifying a person's mental state. Annotation guidelines would need to distinguish personal expressions of distress from quotations, news reports, discussions of another person, and general educational content.

Contextual classification

The NLP component would identify text for further attention based on defined categories and annotated examples. Development would require a documented dataset, an annotation procedure involving suitable domain expertise, and a clear separation between data used to develop the model and data used to evaluate it. Model selection and decision thresholds would be research decisions to examine during implementation.

A later evaluation should report both missed concerning posts and incorrectly flagged posts. Precision, recall, and related measures can describe these errors, while subgroup and language-specific checks can examine whether performance varies across types of text. A numeric classification score should not be presented as a clinical diagnosis or a complete assessment of an individual's circumstances.

Chatbot interface and human support

The chatbot would provide a clear introduction to its role, offer concise support information, and make a route to human assistance visible. Responses should be developed with qualified professionals and checked for relevance, tone, and appropriateness. The interface should let users choose whether to engage and should make it easy to leave the conversation.

A defined referral process would identify which services are available, how information is maintained, and when a human professional should become involved. The design should avoid equating the delivery of an automated response with the successful provision of care. Resource presentation, user engagement, and actual connection to support are distinct outcomes that would need to be examined separately.

Privacy and oversight

Data minimization, access controls, and documented retention rules would be part of the proposed workflow. Reporting should use aggregate results and avoid reproducing identifiable posts without an appropriate basis. Oversight would include review of annotation decisions, response scripts, referral information, and errors observed during development. These requirements are components of the proposed design, rather than findings from the preliminary survey.

Discussion

The preliminary survey points to two practical considerations: respondents reported using several social-media platforms, and most of the five answers to the resource-awareness item indicated no awareness of support resources. Together, these observations support a design that gives users clear information about available help and does not assume that one platform represents their online activity.

The survey findings and the proposed NLP workflow answer different questions. The survey describes reported experiences and awareness. The framework sets out a possible relationship between text processing and support information. The counts do not establish that linguistic classification is accurate, that a chatbot improves help-seeking, or that automated engagement changes mental-health outcomes. Those questions belong to subsequent implementation research.

Several limits affect interpretation. The survey group is small, and the number of tabulated answers varies by question. Recruitment procedures, collection dates, and the reason for the different response totals are not specified in the survey record. The results should therefore be read as preliminary design context. Cross-tabulated counts linking platform use to awareness are not available, so no association between those variables is inferred.

Conclusions

This paper proposes an NLP and chatbot workflow that connects contextual assessment of social-media text with accessible information and referral to human support. The preliminary survey highlights limited awareness of support resources among the five respondents represented in that item and illustrates the use of multiple social-media platforms within the seven-person group.

Further development should begin with a clearly defined use case, appropriate data permissions, domain-informed annotation, and a documented human-referral process. Evaluation should examine classification errors, response quality, usability, privacy, and whether users can reach relevant support. The framework provides a basis for organizing that work while keeping the descriptive survey findings within their intended scope.

References

Arowosegbe, A., & Oyelade, T. (2023). Application of natural language processing (NLP) in detecting and preventing suicide ideation: A systematic review. *International Journal of Environmental Research and Public Health, 20*(2), Article 1514. https://doi.org/10.3390/ijerph20021514

Balcombe, L. (2023). AI chatbots in digital mental health. *Informatics, 10*(4), Article 82. https://doi.org/10.3390/informatics10040082

De Choudhury, M., Kiciman, E., Dredze, M., Coppersmith, G., & Kumar, M. (2016). Discovering shifts to suicidal ideation from mental health content in social media. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 2098–2110). Association for Computing Machinery. https://doi.org/10.1145/2858036.2858207

Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial. *JMIR Mental Health, 4*(2), Article e19. https://doi.org/10.2196/mental.7785

Shing, H.-C., Nair, S., Zirikly, A., Friedenberg, M., Daumé III, H., & Resnik, P. (2018). Expert, crowdsourced, and machine assessment of suicide risk via online postings. *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic* (pp. 25–36). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-0603