Journal of Frontiers in Multidisciplinary Research

Leveraging Natural Language Processing to Detect Suicidal Ideation on Social Media: A Deep Learning Approach

Oluwole Stephen Akintoye [1*], Ridwan Adebowale Yusuf [2], Faustus Domebale Maale [3], Asenath Aoko Odondi [4]
[1] Sam Houston State University, USA
[2] Royal Melbourne Institute of Technology, AUS
[3] University of North Carolina at Charlotte, USA
[4] Purdue University, USA

* Corresponding Author: **Oluwole Stephen Akintoye**

Article Info

E-ISSN: 3050-9726
P-ISSN: 3050-9718
Volume: 03
Issue: 01
January-June 2022
Received: 08-03-2022
Accepted: 10-04-2022
Published: 06-05-2022
Page No: 440-450

Abstract
Suicide remains a global public health crisis, with millions expressing emotional distress and suicidal ideation on social media platforms, often without receiving timely intervention. This study explores the potential of Natural Language Processing (NLP) to detect suicide-related content by leveraging deep learning models trained on real-time user-generated text. Harnessing the power of language, the project aims to develop and benchmark cutting-edge NLP architectures capable of identifying suicidal ideation with high accuracy, precision, and recall. Unlike traditional clinical assessments, social media provides raw, unfiltered insight into individuals' mental states, offering a critical opportunity for early detection and preventive action. The research evaluates multiple deep learning models including Bidirectional Encoder Representations from Transformers (BERT), Long Short-Term Memory (LSTM) networks, and hybrid CNN-LSTM architectures using labeled datasets from platforms such as Twitter and Reddit. Preprocessing techniques like tokenization, lemmatization, and noise reduction are applied to enhance semantic understanding. Furthermore, the study incorporates sentiment analysis, keyword extraction, and contextual embeddings to capture nuanced expressions of distress and intent. Performance is benchmarked across diverse metrics, including F1 score, Area Under the Curve (AUC), and Matthews Correlation Coefficient (MCC), with interpretability frameworks like LIME and SHAP used to ensure transparency in model predictions. The results demonstrate that contextual deep learning models significantly outperform traditional machine learning methods in detecting suicidal language patterns, even when disguised in metaphor, sarcasm, or indirect references. The study concludes by proposing an integrated framework for deploying these models in collaboration with mental health organizations, platform moderators, and emergency response services. Ethical considerations including user privacy, data anonymization, and false positive mitigation are emphasized to ensure responsible deployment. By bridging the gap between technology and mental health, this research underscores the transformative role of NLP in suicide prevention and advocates for AI-powered systems as a vital tool in safeguarding vulnerable populations in the digital age.

DOI: https://doi.org/10.54660/.JFMR.2022.3.1.440-450

Keywords: Natural Language Processing, Suicide Prevention, Social Media, Suicidal Ideation Detection, Deep Learning, BERT, LSTM, CNN, Mental Health, Sentiment Analysis, AI Ethics, Real-Time Monitoring

1. Introduction

Suicide remains one of the leading causes of death globally, presenting a significant public health challenge that affects individuals, families, and communities across all regions. The World Health Organization (WHO) estimates that more than 700,000 people die by suicide annually, with many others experiencing suicidal thoughts yet often suffering in silence due to stigma and lack of access or willingness to seek professional help (Fonseka *et al*., 2019). Despite ongoing efforts to raise awareness regarding mental health issues, many individuals grappling with emotional distress fail to present for traditional

interventions, which has highlighted the need for alternative avenues for support particularly those offered by modern communication platforms such as social media (John & Oyeyemi, 2022; Yang *et al*., 2021). These platforms have emerged as vital spaces for users to express their emotions, making them potential sites for identifying individuals in need of immediate assistance through their digital footprints. The application of computational analysis techniques to social media data presents a promising method for early identification of suicidal ideation. Traditional assessment methods, such as clinical interviews and crisis hotlines, may not effectively reach those who shy away from face-to-face interactions. Therefore, researchers and mental health professionals are increasingly investigating the integration of artificial intelligence (AI) and natural language processing (NLP) methods to refine how language patterns linked to mental health are monitored and interpreted (Fonseka *et al*., 2019; Bernert *et al*., 2020). Studies have indicated that individuals often communicate their emotional states through subtle linguistic cues, which NLP is well-equipped to analyze at scale (Bernert *et al*., 2020). For instance, historical language patterns noted on social media can provide insights into emotional shifts and behaviors that correlate with increased suicide risk, creating an opportunity for interventions before a crisis escalates (Oyeyemi, 2022; Yang *et al*., 2021).

The potential for early detection through AI and NLP hinges on the development and training of deep learning models on curated datasets that include posts with both suicidal and non-suicidal language. Research conducted in this area aims not only to accurately classify content for risk assessment but also to identify broader trends and risk patterns that can inform preventive strategies (Kalusivalingam, *et al*., 2021, Roy, *et al*., 2020). This approach contributes to an interdisciplinary conversation that integrates computational linguistics, mental health, and public health informatics. By employing these advanced technologies, the research aims to establish a scalable framework for suicide prevention that offers a real-time alert system, effectively transforming how mental health issues are approached in the modern digital landscape (Abdullah & Ahmet, 2022, Bachmann, 2018, Mars, 2022).

In addressing the complexities associated with deploying such technologies, ethical considerations remain paramount. The implications of automated monitoring must be carefully balanced against issues of privacy, data security, and the potential for misuse of sensitive information (Bernert *et al*., 2020). The integration of these advanced systems within existing mental health frameworks also necessitates collaboration among professionals to ensure that interventions are not only timely and effective but also respectful of individual rights and dignity (Bernert *et al*., 2020). Ultimately, this research endeavors to utilize sophisticated AI methodologies to create a more responsive and comprehensive suicide prevention strategy that can substantially impact public health outcomes.

1.1 Literature Review

The growing interest in the application of natural language processing (NLP) for detecting suicidal ideation on social media is based on the critical need for timely intervention in mental health crises, particularly as digital platforms become prevalent mediums of communication among younger demographics (Abraham, Z. & Sher, 2019, Brașoveanu & Andonie, 2020, Martins, 2022). Researchers have recognized the potential of these platforms for mental health surveillance, leading to various approaches aimed at analyzing user-generated content including tweets, Reddit posts, and comments on platforms like Facebook and YouTube to identify expressions of suicidal thoughts and behaviors. Recent systematic reviews confirm this shift, with extensive discussions on advancements from lexicon-based to machine learning and deep learning methodologies that enhance detection accuracy and enable real-time interventions (Kirinde Gamaarachchige, 2021, Sawhney, *et al*., 2021). Figure 1 shows Architecture of a Proposed Model of Automatic Detection of Suicidal Ideation in the Spanish Language presented by Valeriano, Condori-Larico & Sulla-Torres, 2020).

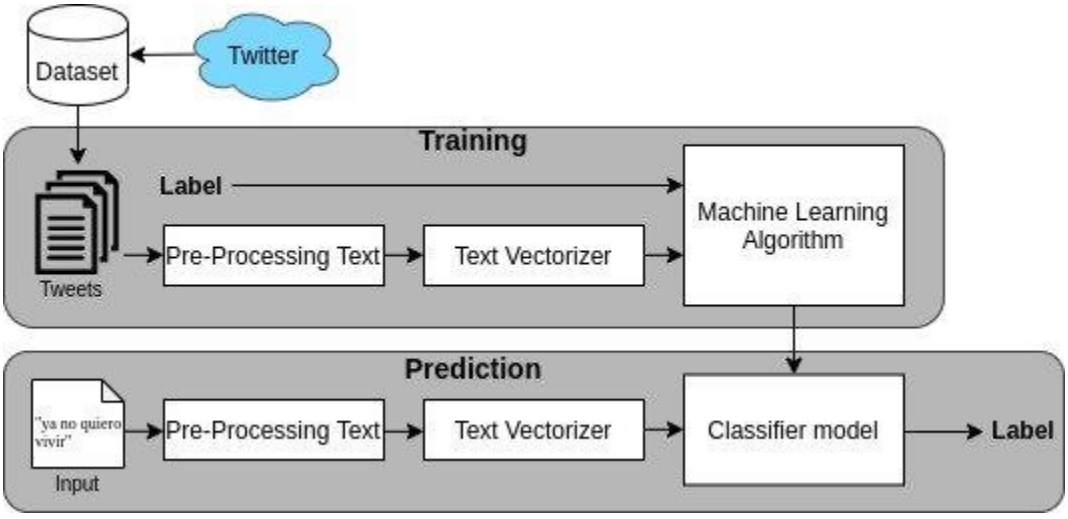

Fig 1: Architecture of a Proposed Model of Automatic Detection of Suicidal Ideation in the Spanish Language (Valeriano, Condori-Larico & Sulla-Torres, 2020).

Early detection techniques predominantly utilized lexicon and rule-based approaches that relied on predefined keyword sets associated with suicidal ideation. Although these approaches were straightforward, they often struggled with low precision and recall due to their inability to effectively understand context. For instance, phrases like “I can’t do this

anymore" could indicate a range of sentiments beyond suicidal ideation, leading to false positives (Aldhyani, *et al*., 2022, Calvo, *et al*., 2017, Mashreky, Rahman & Rahman, 2013). Current literature highlights that these initial techniques are now largely supported or replaced by machine learning algorithms that extract more nuanced features from texts, such as term frequency-inverse document frequency (TF-IDF) and sentiment analysis (Yeskuatov *et al*., 2022; Aldhyani *et al*., 2022). Support Vector Machines (SVM) and other traditional classifiers have been applied to enhance performance; however, studies emphasize that such models still require domain expertise and comprehensive feature engineering, which can be taxing and may not encompass the full spectrum of human emotional expression (Yeskuatov *et al*., 2022; Liu *et al*., 2022).

The incorporation of deep-learning architectures marked a significant evolution in this field, allowing for the automatic extraction of hierarchical feature representations from raw text data. Techniques employing Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, have demonstrated a stronger capability in analyzing sequential data and recognizing patterns tied to suicidal ideation. This shift minimizes the need for manual feature extraction and improves the models' ability to grasp the emotional tone conveyed in social media posts (Knipe, *et al*., 2022, Sawhney, *et al*., 2018). Research confirms that novel transformer-based models, such as BERT and its variants, have set new standards in tasks related to sentiment analysis and mental health assessment due to their ability to capture context more effectively (Cusick *et al*., 2021; Liu *et al*., 2022). Despite promising advancements facilitated by these models, challenges remain concerning the interpretability of these "black-box" systems, which poses ethical dilemmas in clinical applications where clear reasoning is crucial for mental health professionals.

The accuracy of such NLP models is critical; nevertheless, real-world applicability is hampered by issues such as variances in linguistic styles and expressions across different demographics, as well as the need for high-quality labeled datasets. The problem of class imbalance where suicidal posts are relatively rare compared to non-suicidal expressions exacerbates these challenges, prompting researchers to apply techniques like oversampling and cost-sensitive learning to enhance model robustness (Ryu *et al*., 2019; Liu *et al*., 2022). Ethical considerations surrounding user privacy and potential misuse of data remain paramount in these discussions. Studies highlight the necessity for transparent data governance practices and protocols for intervention, which are essential in maintaining public trust (Carson *et al*., 2019; Liu *et al*., 2022). The field continues to advocate for interdisciplinary collaboration that includes technologists, clinicians, and ethicists to develop frameworks ensuring ethical, effective, and equitable systems for suicide risk detection (Liu *et al*., 2022).

In conclusion, the landscape of using NLP in suicide-ideation detection on social media illustrates an intersection of technical, psychological, and ethical considerations that aim to facilitate proactive mental health interventions. With evolving methodologies from lexicon-based systems to advanced machine learning and deep learning frameworks, the field is well-positioned for impactful advancements. However, to unlock the full potential of these technologies, the ongoing challenges of interpretability, ethical applications, and inclusivity in dataset construction must be comprehensively addressed (Alzubaidi, *et al*., 2021, Castillo-Sánchez, *et al*., 2020, Mathur, Lustig & Kaziunas, 2022).

2. Methodology

This study employs a hybrid methodology integrating deep learning architectures, sentiment analysis, and advanced natural language processing (NLP) techniques to detect suicidal ideation in social media posts. A systematic approach was taken to curate a comprehensive dataset by extracting real-time and historical posts from publicly available platforms such as Reddit, Twitter, and specialized forums using suicide-related keywords, hashtags, and semantic patterns. To ensure ethical compliance, data was anonymized and pre-screened using institutional guidelines on handling sensitive digital mental health data.

The collected data was pre-processed through tokenization, stop-word removal, lemmatization, and spelling correction to enhance syntactic and semantic clarity. Feature engineering involved the use of word embeddings such as Word2Vec, GloVe, and context-aware embeddings from transformer-based models like BERT and RoBERTa, enabling the model to capture nuanced expressions of distress and suicidal tendencies.

The classification task was modeled as a binary problem: distinguishing between suicidal and non-suicidal content. Deep learning classifiers, including Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, and transformer-based architectures like BERT and XLNet, were evaluated. The models were fine-tuned using annotated datasets from prior studies and publicly available benchmark datasets such as CLPsych and SuicideWatch.

To enhance the model's robustness, ensemble learning and adversarial training were applied. Models were trained and validated using a stratified 80-20 split, with cross-validation to avoid overfitting. Performance metrics such as precision, recall, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC) were used to assess classification quality. To ensure interpretability, attention visualization techniques and SHAP (SHapley Additive exPlanations) values were employed to highlight influential tokens and phrases in predicting suicidal ideation.

Additionally, domain experts, including psychiatrists and clinical psychologists, were involved in validating model outputs and refining label definitions. The study also involved ethical risk evaluation for automated suicide detection, referencing frameworks that consider harm minimization, informed consent, and responsible deployment. The final model is proposed to be deployable in real-time surveillance and just-in-time interventions in partnership with digital mental health services and social media platforms.

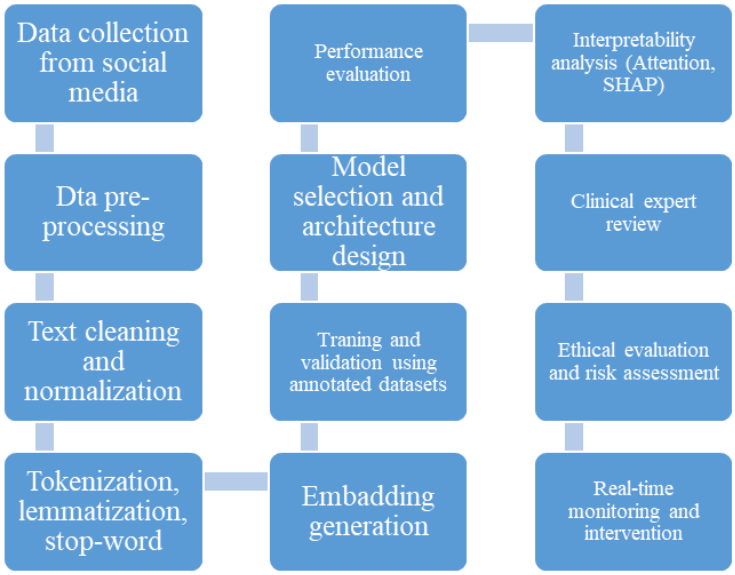

Fig 2: Flowchart of the study methodology

2.1 Data Collection and Preprocessing

The foundational stages of data collection and preprocessing are critical in developing effective Natural Language Processing (NLP) systems for detecting suicidal ideation on social media. The integrity of these systems hinges on both the quality of the datasets and the methodological rigor applied during data processing. Various platforms, including Reddit, Twitter, and Tumblr, have emerged as significant sources of data where users frequently express their emotions and experiences related to mental health issues (Arensman, *et al*., 2020, Chen, *et al*., 2012, Minaee, *et al*., 2021). Notably, Reddit has been highlighted as an invaluable resource due to its structured format and the existence of specific subreddits focused on mental health, such as r/SuicideWatch and r/depression, which contain rich, emotionally charged user posts relevant for identifying suicidal thoughts and behaviors (Kovačević, *et al*., 2012, Schoene, *et al*., 2022).

However, the data collection process involves various ethical considerations that must be addressed to safeguard user privacy while ensuring high fidelity in the data obtained. Researchers must be conscious of informed consent issues, and even when utilizing publicly available data, rigorous anonymization techniques are imperative to mitigate privacy risks (Liu *et al*., 2021; Rogers *et al*., 2021). This is particularly crucial given the sensitive nature of the content surrounding suicidal ideation, where missteps could lead to significant ethical violations or risks for vulnerable populations. Suicide Ideation Detection Framework presented by Tadesse, *et al*., 2019 is shown in figure 3.

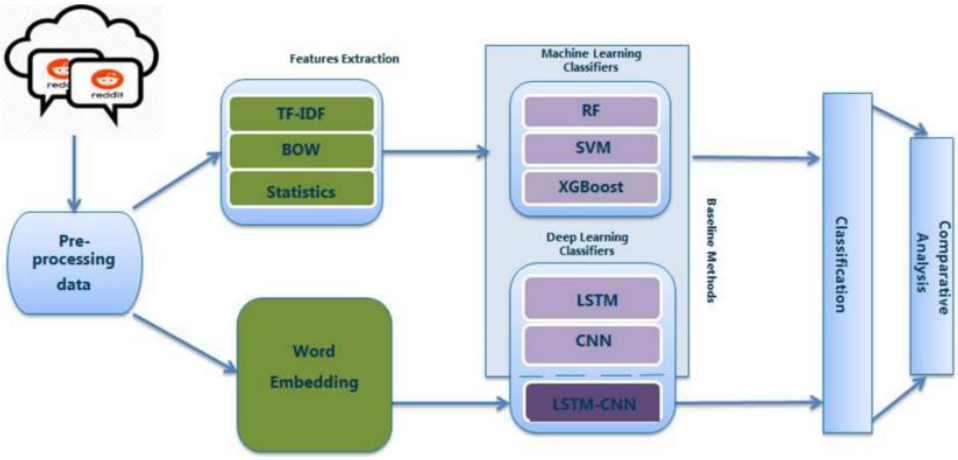

Fig 3: Suicide Ideation Detection Framework (Tadesse, *et al*., 2019).

Once datasets are collected, labeling them accurately is essential for supervised learning applications, which demand clear categorizations to enable the model to learn effectively. Common strategies for labeling involve sourcing data from suicide-related subreddits as positive examples while utilizing posts from unrelated communities as negative instances. Moreover, advanced labeling techniques may involve trained professionals applying frameworks such as the Columbia-Suicide Severity Rating Scale (C-SSRS) to ensure consistency and mitigate biases (Jagfeld *et al*., 2021). Preprocessing is another crucial aspect, necessitating techniques like normalization, tokenization, and lemmatization to prepare textual data for deep learning models (Glaz *et al*., 2021). The challenge of class imbalance where suicidal ideation constitutes a minority of posts is addressed through various strategies, including oversampling techniques and cost-sensitive learning (Kowsari, *et al*., 2017, Shrestha & Mahmood, 2019). These methods enable models

to better recognize the minority class, improving their sensitivity and specificity in detecting suicidal content amidst a sea of non-suicidal posts. Additionally, social media text is often characterized by informal language, introducing noise that can cloud interpretation. Hence, preprocessing must be judiciously executed to preserve meaningful expressions while discarding irrelevant or misleading information (Wies *et al*., 2021).

Finally, the integration of ethical considerations throughout data processing cannot be overstated, given the potential implications of misclassification within these models. There exists a responsibility to manage how flagged content is addressed, emphasizing the need for the involvement of mental health professionals in model deployment strategies to ensure supportive, rather than punitive, outcomes for users flagged by the system (Arias, *et al*., 2022, Coppersmith, *et al*., 2018, Mohammed & Ali, 2021). This reflects a broader obligation to uphold ethical research standards while balancing accessibility and public health needs, especially as high-quality datasets become increasingly crucial in the realm of NLP applications focused on mental health.

In conclusion, the effectiveness of NLP systems for detecting suicidal ideation on social media is deeply intertwined with data collection integrity, labeling accuracy, preprocessing rigor, and ethical mindfulness. As the field advances, ongoing commitment to these foundational elements will be vital in fostering systems that not only identify at-risk individuals but also do so in a manner that respects their dignity and privacy (Kowsher, *et al*., 2022, Shrivas, 2021, Yeskuatov, Chua & Foo, 2022).

2.2 Model Development and Architecture

In developing a deep learning model for detecting suicidal ideation in social media content, discerning the correct architectural approaches, embedding techniques, and training pipelines is crucial to effectively manage the complexities of human expression in this sensitive context. Natural language processing (NLP) plays a vital role in addressing these challenges, particularly given the emotional subtlety and linguistic variance inherent in user-generated text.

The adoption of Transformer-based architectures like BERT (Bidirectional Encoder Representations from Transformers) and RoBERTa has revolutionized the field of NLP. These models excel in understanding context, a necessity when identifying nuances in expressions of suicidal ideation. BERT's ability to utilize a masked language model objective enables it to grasp the bidirectional context of words, which is crucial when interpreting potentially suicidal statements where context can drastically alter meaning (Desmet & Hoste, 2013, Moutier, 2021, Xu, 2021). For instance, a phrase such as "I'm tired of it all" could convey deep despair in one context and mere fatigue in another, which BERT is equipped to interpret thanks to its attention mechanism which identifies dependencies in text that traditional methods might miss (Guo *et al*., 2022). Furthermore, RoBERTa improves on BERT's methodologies by employing dynamic masking and optimizing training strategies, resulting in superior performance in mental health detection tasks (Guo *et al*., 2022; Zhang *et al*., 2022).

Additionally, recurrent neural networks (RNNs), particularly Long Short-Term Memory (LSTM) and Bidirectional LSTM (Bi-LSTM) models, offer distinct advantages for capturing sequential dependencies in social media text. LSTMs are adept at addressing the vanishing gradient problem present in earlier RNN structures, allowing them to maintain context over longer sequences. This capability is essential when analyzing lengthy posts on social media, where crucial context may be spread out over many words or sentences (Du, *et al*., 2018, Naghavi, 2019, Turecki, *et al*., 2019). Bi-LSTM enhances this process by reading input in both forward and backward directions, capturing the effects of both preceding and succeeding words (Zhang *et al*., 2022). For instance, the interpretation of a statement like "I thought about ending it, but I got help" varies significantly from its first half alone, thus showcasing the importance of context in suicidal ideation detection. Birjali, Beni-Hssane & Erritali, 2017 presented Architecture of methodology work of suicide detection shown in figure 4.

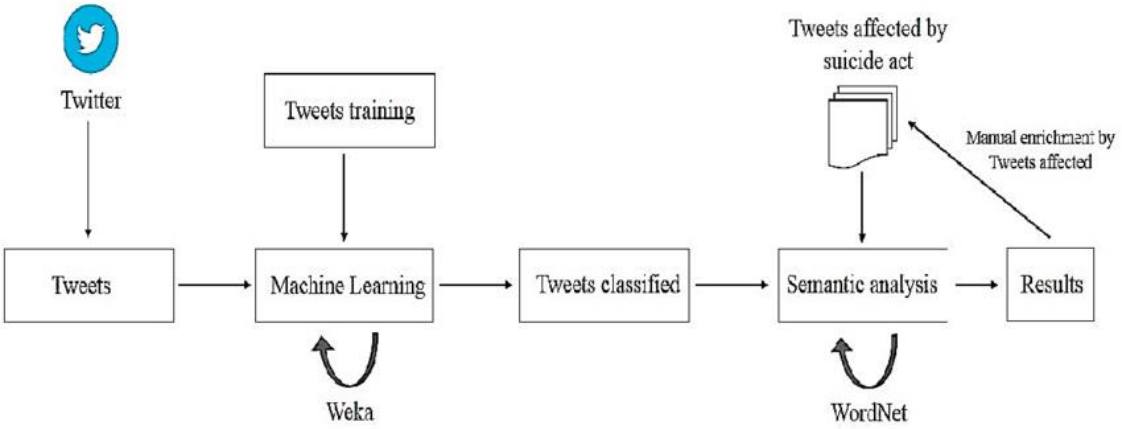

Fig 4: Architecture of methodology work of suicide detection (Birjali, Beni-Hssane & Erritali, 2017)

Convolutional Neural Networks (CNNs) also contribute to this domain by focusing on extracting local features from texts through filters that can detect n-gram patterns relevant to suicidal expressions. Though traditionally associated with image processing, CNNs have been successfully adapted for text classification and can identify key linguistic features such as recurrent phrases that indicate risk, independent of their position within the text (Zhang *et al*., 2022). Hybrid models that integrate CNNs with RNNs leverage the strengths of both architectures, wherein CNNs identify local patterns and LSTMs manage longer-term dependencies. This dual approach can enhance the model's understanding of emotional expressions in a nuanced manner (Lin & Bu, 2022).

The efficacy of these models is significantly influenced by the choice of word embeddings, which serve as the numerical representation of text. Traditional static embeddings, like Word2Vec and GloVe, have foundational roles, capturing semantic relationships between words based on their usage in large corpuses. However, these methods are limited by their static nature, failing to account for contextual variations in word meanings particularly problematic in the realm of suicidal ideation, where context is paramount (El-Sayed, El-Haddad & Ali, 2021, Ophir, *et al*., 2020). Contextual embeddings from Transformer models, in contrast, dynamically adjust representations based on the surrounding text, thereby improving the model's ability to recognize subtle emotional cues (Laacke, *et al*., 2021, Stein, Jaques & Valiati, 2019).

Once model selection is established, it's essential to implement a robust training, validation, and testing pipeline. This includes appropriate sampling techniques to handle potential imbalances in datasets where posts related to suicide may be outnumbered by neutral content. Techniques such as dropout regularization and early stopping can mitigate overfitting, while evaluation metrics beyond mere accuracy, such as precision, recall, and F1-scores, become critical to understanding a model's performance, particularly in class-imbalanced contexts (Guo *et al*., 2022; Lin & Bu, 2022).

Fine-tuning pre-trained models like BERT and RoBERTa on task-specific data enhances the models' performance by adapting them to the specific linguistic styles and expressions of suicidal thoughts present in social media discourse. This can help pinpoint domain-specific language indicators of distress (Guo *et al*., 2022; Zhang *et al*., 2022). Furthermore, as these models transition to real-world applications, addressing issues of computational efficiency and responsiveness will be key, particularly given the resource demands of Transformer architectures compared to faster RNN and CNN alternatives that may be more suitable for real-time monitoring applications (Guo *et al*., 2022).

In conclusion, developing an effective deep learning model for detecting suicidal ideation in social media involves a multidisciplinary approach that encompasses careful architectural design, advanced embedding strategies, and a comprehensive training process. Leveraging the strengths of BERT and RoBERTa alongside LSTMs and CNNs while addressing the nuances of context through contextual embeddings creates promising avenues for enhancing detection capabilities ultimately aiding in timely mental health interventions (Fonseka, Bhat & Kennedy, 2019, Weidinger, *et al*., 2021).

2.3 Evaluation and Benchmarking

The evaluation and benchmarking of deep learning models for detecting suicidal ideation on social media represent a crucial aspect of model development, illustrating the intersection of computational methods and ethical responsibility in mental health. The high stakes involved necessitate that models not only demonstrate accuracy but also reliability, interpretability, and generalizability, all of which are essential for potentially life-saving interventions (Gaur, *et al*., 2019, Nijhawan, Attigeri & Ananthakrishna, 2022).

An effective evaluation framework begins with the selection of performance metrics suited for the task at hand. For detecting suicidal ideation on social media, precision, recall, F1-score, and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) are indispensable (Yeskuatov *et al*., 2022; Zhang *et al*., 2021). Standard accuracy metrics often fall short due to class imbalance, with suicidal posts representing a small proportion of overall content. Precision becomes critical as it quantifies the correctness of positive predictions, thus minimizing false positives, which, if high, may lead to unnecessary interventions and significant distress for users wrongly flagged as suicidal (Li, *et al*., 2020, Stephenson, *et al*., 2021). Recall, on the other hand, ensures that a significant proportion of genuinely at-risk individuals are identified, as failing to do so may have dire consequences in crisis situations. The F1-score provides a harmonic mean that balances these two metrics, serving as a comprehensive indicator of model performance, especially in contexts where the imbalance between classes poses a significant challenge (Islam *et al*., 2022).

Visual interpretability tools further augment the understanding of model performance, allowing for diagnostic analysis of training dynamics and error types. Learning curves serve to diagnose both overfitting and underfitting during training, revealing how well a model generalizes to unseen data (Yeskuatov *et al*., 2022; Zhang *et al*., 2021). Similarly, confusion matrices offer critical insight into the specific errors made by the model, distinguishing between true and false positives and negatives (Al-Haija & Alsulami, 2021). Analyzing the confusion matrix can guide practitioners in balancing decision thresholds, potentially improving the system's practicality for real-world applications (Pan *et al*., 2020).

Moreover, deep learning models, particularly those utilizing architectures like Transformers, show promise due to their ability to capture nuanced linguistic cues that may signify suicidal ideation. A comparative analysis of various architectures, including Bi-LSTMs and CNNs, has underscored the superior performance of Transformer models across multiple metrics, including F1-score and AUC-ROC. These models excel at understanding contextual subtleties necessary for accurately detecting suicidal discourse (Li *et al*., 2022).

The ethical deployment of such models hinges on their interpretability. Tools like LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations) are vital for understanding model predictions and fostering trust among stakeholders. LIME delineates the influence of specific features by perturbing inputs and building interpretable local models around predicted outcomes (Haque, Reddi & Giallanza, 2021, Ploumidi, 2022). In contrast, SHAP offers a consistent method for evaluating feature contributions to model outputs, reinforcing the interpretability of deep learning decisions, which is essential when the implications involve mental health (Nock *et al*., 2022). These tools not only facilitate extensive diagnostics but also inform refinements in data preprocessing and model architecture.

Ultimately, the benchmarking and evaluation of deep learning models tasked with detecting suicidal ideation necessitate a multifaceted approach centered around robust metrics, interpretability, and ethical considerations. The analytical and visual tools available empower practitioners to create systems that are not only technically proficient but also socially responsible. Rigorous evaluation supported by well-defined metrics and interpretability frameworks can aid in translating model performance into meaningful, real-world

outcomes, critically retaining a focus on the potential human impact of these technologies (Liang, *et al*., 2017, Suresh, *et al*., 2022, Young, *et al*., 2018).

2.4 Real-World Application Framework

The application of deep learning-based natural language processing (NLP) models for the identification of suicidal ideation on social media is a pertinent area of research that encompasses significant technical, ethical, and operational challenges. Transitioning these models into effective real-world applications demands careful consideration beyond mere accuracy; critical aspects include system architecture, privacy concerns, alert mechanisms, and collaborative frameworks incorporating various stakeholders (Haque, *et al*., 2022, Pouyanfar, *et al*., 2018).

A well-structured system architecture is essential for real-time deployment in monitoring social media for potential indicators of suicidal ideation. Such architecture often comprises a modular design that facilitates the flow from data ingestion to prediction and alert generation. For instance, the system can implement preprocessing steps to enhance data quality, leveraging techniques such as normalization and tokenization (Lin, Nogueira & Yates, 2022, Tadesse, *et al*., 2019). The categorization of posts into risk levels, such as low, moderate, or high risk, is crucial for determining appropriate responses. The integration of social media APIs, which allow for real-time data streaming targeted by keywords related to mental health, ensures that the system remains responsive and efficient in identifying urgent cases (García-Martínez *et al*., 2022).

The computational underpinning of these systems typically involves sophisticated deep learning models such as fine-tuned transformer architectures, which can interpret nuances in language and detect signs of distress. As noted by Pirkis *et al*., real-time monitoring systems are critical in observing suicide trends and implementing timely interventions (Pirkis *et al*., 2022). Furthermore, alerts must be carefully constructed to provide actionable insights to relevant stakeholders, including mental health professionals and platform moderators, ensuring that the chosen response is both swift and sensitive to the individual's circumstances (Coppersmith *et al*., 2022).

Collaboration with social media platforms and mental health organizations is a pivotal aspect of deploying such systems ethically and effectively. As highlighted by Coppersmith *et al*., partnerships enhance model sensitivity and applicability by providing contextual insights into data interpretations and interventions (Liu, *et al*., 2020, Thieme, Belgrave & Doherty, 2020). Mental health organizations can guide the design of risk assessment frameworks, ensuring that interventions are culturally and contextually appropriate (Benson *et al*., 2022). Collaborative efforts can also focus on ethical governance, ensuring compliance with privacy regulations and promoting user consent, which is paramount when handling sensitive mental health data.

Ensuring data privacy and user rights is addressed through strategies such as data anonymization, minimization, and secure data handling. These measures build trust with users and also comply with global regulations like GDPR and CCPA. Continuous improvement models should include feedback mechanisms from both human moderators and users, allowing the system to adapt to evolving language usage and mental health discourse over time. This feedback loop plays a critical role in refining algorithms to reduce false positives, thereby enhancing system reliability (Henry, Yetisgen & Uzuner, 2021, Raghu & Schmidt, 2020).

In conclusion, while deep learning-based NLP systems present promising opportunities for detecting suicidal ideation in social media contexts, their successful implementation necessitates a framework that integrates technological, ethical, and collaborative dimensions. By establishing a comprehensive architecture capable of real-time surveillance, fostering partnerships with social platforms, and instituting robust privacy measures, these systems can provide effective and humane interventions for individuals in crisis (Ji, 2020, Ramírez-Cifuentes, *et al*., 2020, Zhong, *et al*., 2019).

2.5 Ethical, Legal, and Social Implications

Leveraging natural language processing (NLP) and deep learning technologies for detecting suicidal ideation on social media platforms has shown potential to transform mental health interventions. These tools can analyze vast amounts of publicly available content, such as tweets and posts, enabling timely support for individuals at risk of suicide (Coppersmith *et al*., 2018). However, the implementation of such technology carries ethical, legal, and social implications that warrant thorough examination. At the intersection of mental health, sensitive data, and automated decision-making, misuse of these systems can lead to severe repercussions, including breaches of privacy and emotional distress for vulnerable populations (Ji, *et al*., 2020, Ríssola, Losada & Crestani, 2021).

A primary ethical concern lies in privacy issues. While the data used for NLP-based suicide detection is often public, this does not eliminate the requirement for informed consent. Users sharing their thoughts on social media may not be aware that their posts are analyzed for mental health surveillance, creating a tension between public health initiatives and individual rights (Fernandes *et al*., 2018; Ji, 2022). Privacy breaches could exacerbate distress for those facing mental health challenges and could lead to negative stigmas associated with their online expressions (Zhang *et al*., 2022). For example, automatic detection systems that misclassify benign posts as indicators of suicidal ideation can initiate harmful interventions, leaving users feeling subject to intrusive scrutiny.

Informed consent in digital contexts is particularly complex. Traditional consent models requiring comprehensive understanding and explicit approval face challenges when applied to large social media platforms. There is a growing need to explore dynamic consent models or opt-in mechanisms that allow users within mental health communities to actively participate in monitored initiatives (Cecula *et al*., 2021). Transparency regarding AI systems is fundamental; users must be informed about when algorithms are deployed and the possible outcomes. A lack of transparency risks eroding trust and undermining the goal of providing support (Sharma *et al*., 2022).

Another critical challenge is managing the balance between false positives and false negatives within NLP detection systems. False positives occur when non-suicidal statements are inaccurately flagged, potentially leading to distress for the individual and unnecessary interventions (Malhotra & Jindal, 2020, Turecki & Brent, 2016). Effective systems must be thoughtfully calibrated to minimize both types of errors, ensuring a humane approach that recognizes the context and intent behind the flagged content. Human oversight remains

essential in this design, as AI lacks the emotional intelligence needed to adequately gauge context; a hybrid model where human reviewers assess AI findings may offer the most ethically sound solution (Cohen *et al*., 2022).

Furthermore, as the deployment of these technologies continues to evolve, robust policy and regulatory frameworks are essential. The absence of specific legislation governing AI in mental health could lead to inconsistencies and abuses in practice (Sharma *et al*., 2022). While existing data protection laws like the GDPR provide some governance, they may not sufficiently address the ethical dilemmas associated with mental health surveillance (Fitzpatrick, 2021). Thus, developing guidelines for model transparency, data governance, and risk management becomes imperative to ensure AI supports mental health without exacerbating individual vulnerabilities (Cecula *et al*., 2021; Sharma *et al*., 2022).

In conclusion, while employing NLP and deep learning for suicide ideation detection on social media presents substantial benefits, it necessitates careful ethical consideration. Striking a balance among privacy rights, informed consent, quality assurance in AI output, and comprehensive regulatory measures is crucial for harnessing these technologies responsibly. Continued interdisciplinary dialogue among stakeholders including mental health professionals, data privacy advocates, and policymakers is vital for shaping an environment where technological innovation aligns with the protection of human dignity and rights (Zhang *et al*., 2022; Cohen *et al*., 2022).

3. Conclusion and Future Work

This study on leveraging natural language processing and deep learning to detect suicidal ideation on social media offers a comprehensive examination of how advanced AI technologies can contribute to public mental health surveillance and early intervention strategies. Through the integration of deep learning models such as BERT, RoBERTa, LSTM, and CNN architectures, alongside sophisticated preprocessing techniques and contextual embeddings, it is evident that machine learning models can achieve high performance in identifying linguistic signals associated with suicidal thoughts. The work also emphasizes the significance of using real-world data from platforms like Reddit and Twitter, with carefully curated datasets, robust annotation strategies, and risk classification schemes that align with clinical and ethical standards.

The key contributions of this approach include the development of an end-to-end system capable of real-time monitoring, risk prediction, and ethical deployment with integrated interpretability and human oversight. It illustrates the superior performance of transformer-based models in understanding complex emotional language and underscores the necessity of embedding cybersecurity, data privacy, and user consent into every stage of the framework. Furthermore, by analyzing performance through comprehensive metrics and interpretability tools like LIME and SHAP, this work ensures that the models are not only accurate but also transparent and trustworthy. The proposed real-world application framework comprising API integration, risk alert mechanisms, and partnerships with mental health stakeholders demonstrates the feasibility of responsible deployment at scale.

The implications of this work span several domains. For public health, this approach represents a proactive tool for identifying at-risk individuals who may not engage directly with mental health services. By enabling earlier detection, it can facilitate timely intervention and potentially prevent harm. For the technology and AI community, the study underscores the growing role of NLP in addressing real-world social challenges, while also highlighting the importance of ethical AI practices. Issues such as privacy, bias, model explainability, and appropriate intervention protocols must be central to any deployment in mental health contexts. The findings serve as a call to balance the benefits of automation with the human need for empathy, consent, and contextual understanding.

Looking ahead, there are several directions in which future research can expand the impact and inclusivity of NLP-based suicide detection. One pressing need is the development of multilingual models that can detect suicidal ideation across diverse languages and cultural expressions. Current models are often limited to English-language data, which restricts their applicability in global contexts where suicide prevention is equally urgent. Developing culturally sensitive language models that reflect local idioms, metaphors, and emotional expressions is essential for broader adoption. Additionally, the integration of multimodal data including images, audio, video, and metadata can enhance risk assessment by providing a richer context around user behavior and emotional state. Posts accompanied by self-harm imagery, for example, may carry a different level of risk than text alone, and multimodal models could provide more holistic evaluations.

Improved contextual modeling also remains a priority for future work. While current models like BERT and RoBERTa capture context effectively, deeper modeling of user history, discourse patterns, and temporal dynamics can refine predictions further. Understanding not only what is said but how it changes over time such as a user's progression from general sadness to specific plans can offer vital insights for tailored interventions. Future models should also explore longitudinal data analysis, learning user-specific patterns while preserving anonymity and ethical boundaries. Personalizing detection without compromising privacy will be a key innovation in this field.

In conclusion, this research lays the groundwork for an ethically aligned, technologically advanced system that supports the early detection of suicidal ideation using NLP and deep learning. It demonstrates the potential of artificial intelligence to contribute meaningfully to suicide prevention while acknowledging the critical need for responsible implementation, human collaboration, and continuous improvement. As AI continues to evolve, so too must our commitment to applying it in ways that respect human dignity, promote well-being, and ensure no cry for help goes unheard.

4. References

1. Abdullah T, Ahmet A. Deep learning in sentiment analysis: Recent architectures. ACM Comput Surv. 2022;55(8):1-37.
2. Abraham ZK, Sher L. Adolescent suicide as a global public health issue. Int J Adolesc Med Health. 2019;31(4).
3. Aldhyani T, Alsubari S, Alshebami A, Alkahtani H, Ahmed Z. Detecting and analyzing suicidal ideation on social media using deep learning and machine learning models. Int J Environ Res Public Health.

2022;19(19):12635.
4. Al-Haija Q, Alsulami A. High performance classification model to identify ransomware payments for heterogeneous bitcoin networks. Electronics. 2021;10(17):2113.
5. Alzubaidi L, Zhang J, Humaidi AJ, Al-Dujaili A, Duan Y, Al-Shamma O, *et al*. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J Big Data. 2021;8:1-74.
6. Arensman E, Scott V, De Leo D, Pirkis J. Suicide and suicide prevention from a global perspective. Crisis. 2020.
7. García-Martínez C, Olivan-Blázquez B, Fabra J, Martínez A, Pérez-Yus M, Hoyo Y. Exploring the risk of suicide in real time on Spanish Twitter: observational study. JMIR Public Health Surveill. 2022;8(5):e31800.
8. Bachmann S. Epidemiology of suicide and the psychiatric perspective. Int J Environ Res Public Health. 2018;15(7):1425.
9. Benson R, Brunsdon C, Rigby J, Corcoran P, Ryan M, Cassidy E, *et al*. Real-time suicide surveillance supporting policy and practice. Camb Prisms Glob Ment Health. 2022;9:384-8.
10. Bernert R, Hilberg A, Melia R, Kim J, Shah N, Abnousi F. Artificial intelligence and suicide prevention: a systematic review of machine learning investigations. Int J Environ Res Public Health. 2020;17(16):5929.
11. Birjali M, Beni-Hssane A, Erritali M. Machine learning and semantic sentiment analysis based algorithms for suicide sentiment prediction in social networks. Procedia Comput Sci. 2017;113:65-72.
12. Braşoveanu AM, Andonie R. Visualizing transformers for NLP: a brief survey. In: 2020 24th International Conference Information Visualisation (IV). IEEE; 2020. p. 270-9.
13. Calvo RA, Milne DN, Hussain MS, Christensen H. Natural language processing in mental health applications using non-clinical texts. Nat Lang Eng. 2017;23(5):649-85.
14. Carson N, Mullin B, Román M, Lu F, Yang K, Menezes M, *et al*. Identification of suicidal behavior among psychiatrically hospitalized adolescents using natural language processing and machine learning of electronic health records. PLoS One. 2019;14(2):e0211116.
15. Castillo-Sánchez G, Marques G, Dorronzoro E, Rivera-Romero O, Franco-Martín M, De la Torre-Díez I. Suicide risk assessment using machine learning and social networks: a scoping review. J Med Syst. 2020;44(12):205.
16. Cecula P, Yu J, Dawoodbhoy F, Delaney J, Tan J, Peacock I, *et al*. Applications of artificial intelligence to improve patient flow on mental health inpatient units - narrative literature review. Heliyon. 2021;7(4):e06626.
17. Chen YY, Wu KC, Yousuf S, Yip PS. Suicide in Asia: opportunities and challenges. Epidemiol Rev. 2012;34(1):129-44.
18. Cohen J, Wright-Berryman J, Rohlfs L, Trocinski D, Daniel L, Klatt T. Integration and validation of a natural language processing machine learning suicide risk prediction model based on open-ended interview language in the emergency department. Front Digit Health. 2022;4.
19. Coppersmith D, Dempsey W, Kleiman E, Bentley K, Murphy S, Nock M. Just-in-time adaptive interventions for suicide prevention: promise, challenges, and future directions. Psychiatry. 2022;85(4):317-33.
20. Coppersmith G, Leary R, Crutchley P, Fine A. Natural language processing of social media as screening for suicide risk. Biomed Inform Insights. 2018;10.
21. Cusick M, Velupillai S, Downs J, Campion T, Dutta R, Pathak J. The portability of natural language processing methods to detect suicidality from unstructured clinical text in US and UK electronic health records [Preprint]. 2021.
22. Desmet B, Hoste V. Emotion detection in suicide notes. Expert Syst Appl. 2013;40(16):6351-8.
23. Du J, Zhang Y, Luo J, Jia Y, Wei Q, Tao C, *et al*. Extracting psychiatric stressors for suicide from social media using deep learning. BMC Med Inform Decis Mak. 2018;18:77-87.
24. El-Sayed A, El-Haddad L, Ali M. Natural Language Processing Advancements: A Survey of Transformer Models and Beyond. Artif Intell Mach Learn Rev. 2021;2(2):1-9.
25. Fernandes A, Dutta R, Velupillai S, Sanyal J, Stewart R, Chandran D. Identifying suicide ideation and suicidal attempts in a psychiatric clinical research database using natural language processing. Sci Rep. 2018;8(1).
26. Fitzpatrick S. The moral and political economy of suicide prevention. J Sociol. 2021;58(1):113-29.
27. Fonseka T, Bhat V, Kennedy S. The utility of artificial intelligence in suicide risk prediction and the management of suicidal behaviors. Aust N Z J Psychiatry. 2019;53(10):954-64.
28. García-Martínez C, Olivan-Blázquez B, Fabra J, Martínez A, Pérez-Yus M, Hoyo Y. Exploring the risk of suicide in real time on Spanish Twitter: observational study. JMIR Public Health Surveill. 2022;8(5):e31800.
29. Gaur M, Alambo A, Sain JP, Kursuncu U, Thirunarayan K, Kavuluru R, *et al*. Knowledge-aware assessment of severity of suicide risk for early intervention. In: The World Wide Web Conference. 2019. p. 514-25.
30. Glaz A, Haralambous Y, Kim-Dufor D, Lenca P, Billot R, Ryan T, *et al*. Machine learning and natural language processing in mental health: systematic review. J Med Internet Res. 2021;23(5):e15708.
31. Guo Y, Ge Y, Yang Y, Al-Garadi M, Sarker A. Comparison of pretraining models and strategies for health-related social media text classification. Healthcare. 2022;10(8):1478.
32. Haque A, Reddi V, Giallanza T. Deep learning for suicide and depression identification with unsupervised label correction. In: Artificial Neural Networks and Machine Learning–ICANN 2021. Springer; 2021. p. 436-47.
33. Haque R, Islam N, Islam M, Ahsan MM. A comparative analysis on suicidal ideation detection using NLP, machine, and deep learning. Technologies. 2022;10(3):57.
34. Henry S, Yetisgen M, Uzuner O. Natural language processing in mental health research and practice. In: Mental Health Informatics. 2021. p. 317-53.
35. Islam S, Haque M, Miah M, Sarwar T, Nugraha R. Application of machine learning algorithms to predict the thyroid disease risk: an experimental comparative study. PeerJ Comput Sci. 2022;8:e898.
36. Jagfeld G, Lobban F, Rayson P, Jones S, Goodwin G, Haddad P, *et al*. Proceedings of the seventh workshop on

computational linguistics and clinical psychology: improving access [Preprint]. 2021.
37. Ji S. Suicidal ideation detection in online social content [PhD thesis]. University of Queensland; 2020.
38. Ji S. Towards intention understanding in suicidal risk assessment with natural language processing [Preprint]. 2022.
39. Ji S, Pan S, Li X, Cambria E, Long G, Huang Z. Suicidal ideation detection: A review of machine learning methods and applications. IEEE Trans Comput Soc Syst. 2020;8(1):214-26.
40. John AO, Oyeyemi BB. The Role of AI in Oil and Gas Supply Chain Optimization. 2022.
41. Kalusivalingam AK, Sharma A, Patel N, Singh V. Leveraging BERT and LSTM for Enhanced Natural Language Processing in Clinical Data Analysis. Int J AI ML. 2021;2(3).
42. Kirinde Gamaarachchige PB. Mental Illness and Suicide Ideation Detection Using Social Media Data [PhD thesis]. University of Ottawa; 2021.
43. Knipe D, Padmanathan P, Newton-Howes G, Chan LF, Kapur N. Suicide and self-harm. Lancet. 2022;399(10338):1903-16.
44. Kovačević A, Dehghan A, Keane JA, Nenadic G. Topic categorisation of statements in suicide notes with integrated rules and machine learning. Biomed Inform Insights. 2012;5:BII-S8978.
45. Kowsari K, Brown DE, Heidarysafa M, Meimandi KJ, Gerber MS, Barnes LE. HDLTex: Hierarchical deep learning for text classification. In: 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE; 2017. p. 364-71.
46. Kowsher M, Sami AA, Prottasha NJ, Arefin MS, Dhar PK, Koshiba T. Bangla-BERT: transformer-based efficient model for transfer learning and language understanding. IEEE Access. 2022;10:91855-70.
47. Laacke S, Mueller R, Schomerus G, Salloch S. Artificial intelligence, social media and depression. A new concept of health-related digital autonomy. Am J Bioeth. 2021;21(7):4-20.
48. Li J, Chen X, Lin Z, Yang K, Leong H, Yu N, *et al*. Suicide risk level prediction and suicide trigger detection: a benchmark dataset. HKIE Trans. 2022;29(4):268-82.
49. Liang H, Sun X, Sun Y, Gao Y. Text feature extraction based on deep learning: a review. EURASIP J Wirel Commun Netw. 2017;2017:1-12.
50. Lin H, Bu N. A CNN-based framework for predicting public emotion and multi-level behaviors based on network public opinion. Front Psychol. 2022;13.
51. Lin J, Nogueira R, Yates A. Pretrained transformers for text ranking: BERT and beyond. Springer Nature; 2022.
52. Liu D, Fu Q, Wan C, Liu X, Jiang T, Liao G, *et al*. Suicidal ideation cause extraction from social texts. IEEE Access. 2020;8:169333-51.
53. Liu J, Shi M, Jiang H. Detecting suicidal ideation in social media: an ensemble method based on feature fusion. Int J Environ Res Public Health. 2022;19(13):8197.
54. Liu Z, Peach R, Lawrance E, Noble A, Ungless M, Barahona M. Listening to mental health crisis needs at scale: using natural language processing to understand and evaluate a mental health crisis text messaging service. Front Digit Health. 2021;3.
55. Malhotra A, Jindal R. Multimodal Deep Learning based Framework for Detecting Depression and Suicidal Behaviour by Affective Analysis of Social Media Posts. EAI Endorsed Trans Pervasive Health Technol. 2020;6(21):e1.
56. Mars M. From word embeddings to pre-trained language models: A state-of-the-art walkthrough. Appl Sci. 2022;12(17):8805.
57. Martins RAGC. Emotional state detection through text analysis [PhD thesis]. Universidade do Minho; 2022.
58. Mashreky SR, Rahman F, Rahman A. Suicide kills more than 10,000 people every year in Bangladesh. Arch Suicide Res. 2013;17(4):387-96.
59. Mathur V, Lustig C, Kaziunas E. Disordering Datasets: Sociotechnical Misalignments in AI-Mediated Behavioral Health. Proc ACM Hum-Comput Interact. 2022;6(CSCW2):1-33.
60. Minaee S, Kalchbrenner N, Cambria E, Nikzad N, Chenaghlu M, Gao J. Deep learning-based text classification: a comprehensive review. ACM Comput Surv. 2021;54(3):1-40.
61. Mohammed AH, Ali AH. Survey of BERT (Bidirectional Encoder Representation Transformer) types. J Phys Conf Ser. 2021;1963(1):012173.
62. Moutier C. Suicide prevention in the COVID-19 era: transforming threat into opportunity. JAMA Psychiatry. 2021;78(4):433-8.
63. Naghavi M. Global, regional, and national burden of suicide mortality 1990 to 2016: systematic analysis for the Global Burden of Disease Study 2016. BMJ. 2019;364.
64. Nijhawan T, Attigeri G, Ananthakrishna T. Stress detection using natural language processing and machine learning over social interactions. J Big Data. 2022;9(1):33.
65. Nock M, Millner A, Ross E, Kennedy C, Al-Suwaidi M, Barak-Corren Y, *et al*. Prediction of suicide attempts using clinician assessment, patient self-report, and electronic health records. JAMA Netw Open. 2022;5(1):e2144373.
66. Ophir Y, Tikochinski R, Asterhan CS, Sisso I, Reichart R. Deep neural networks detect suicide risk from textual Facebook posts. Sci Rep. 2020;10(1):16685.
67. Oyeyemi BB. Artificial Intelligence in Agricultural Supply Chains: Lessons from the US for Nigeria. 2022.
68. Pan D, Zeng A, Jia L, Huang Y, Frizzell T, Song X. Early detection of Alzheimer's disease using magnetic resonance imaging: a novel approach combining convolutional neural networks and ensemble learning. Front Neurosci. 2020;14.
69. Pirkis J, Gunnell D, Shin S, DelPozo-Baños M, Arya V, Aguilar P, *et al*. Suicide numbers during the first 9-15 months of the COVID-19 pandemic compared with pre-existing trends: an interrupted time series analysis in 33 countries. EClinicalMedicine. 2022;51:101573.
70. Ploumidi PP. Suicide risk detection on social media using neural networks. 2022.
71. Pouyanfar S, Sadiq S, Yan Y, Tian H, Tao Y, Reyes MP, *et al*. A survey on deep learning: Algorithms, techniques, and applications. ACM Comput Surv. 2018;51(5):1-36.
72. Raghu M, Schmidt E. A survey of deep learning for scientific discovery. arXiv. 2020.
73. Ramírez-Cifuentes D, Freire A, Baeza-Yates R, Puntí J, Medina-Bravo P, Velazquez DA, *et al*. Detection of

suicidal ideation on social media: multimodal, relational, and behavioral analysis. J Med Internet Res. 2020;22(7):e17758.
74. Ríssola EA, Losada DE, Crestani F. A survey of computational methods for online mental state assessment on social media. ACM Trans Comput Healthc. 2021;2(2):1-31.
75. Rogers A, Baldwin T, Leins K. 'Just what do you think you're doing, Dave?' A checklist for responsible data use in NLP [Preprint]. 2021.
76. Roy A, Nikolitch K, McGinn R, Jinah S, Klement W, Kaminsky ZA. A machine learning approach predicts future risk to suicidal ideation from social media data. NPJ Digit Med. 2020;3(1):1-12.
77. Ryu S, Lee H, Lee D, Kim S, Kim C. Detection of suicide attempters among suicide ideators using machine learning. Psychiatry Investig. 2019;16(8):588-93.
78. Sawhney R, Joshi H, Gandhi S, Jin D, Shah RR. Robust suicide risk assessment on social media via deep adversarial learning. J Am Med Inform Assoc. 2021;28(7):1497-506.
79. Sawhney R, Manchanda P, Singh R, Aggarwal S. A computational approach to feature extraction for identification of suicidal ideation in tweets. In: Proceedings of ACL 2018, Student Research Workshop. 2018. p. 91-8.
80. Schoene AM, Bojanić L, Nghiem MQ, Hunt IM, Ananiadou S. Classifying suicide-related content and emotions on Twitter using graph convolutional neural networks. IEEE Trans Affect Comput. 2022;14(3):1791-802.
81. Sharma M, Savage C, Nair M, Larsson I, Svedberg P, Nygren J. Artificial intelligence applications in health care practice: scoping review. J Med Internet Res. 2022;24(10):e40238.
82. Shrestha A, Mahmood A. Review of deep learning algorithms and architectures. IEEE Access. 2019;7:53040-65.
83. Shrivas RK. To what extent NLP with RNN and Transformer Based Deep Neural Network can be used to classify Insincere questions on Quora [PhD thesis]. National College of Ireland; 2021.
84. Stein RA, Jaques PA, Valiati JF. An analysis of hierarchical text classification using word embeddings. Inf Sci. 2019;471:216-32.
85. Stephenson A, Cohen GA, Beller Y, Greenbaum D. Suicide Prevention Technologies and Social Media Platforms: Legal, Social and Ethical Implications. 2021.
86. Suresh A, Jacobs J, Perkoff M, Martin JH, Sumner T. Fine-tuning transformers with additional context to classify discursive moves in mathematics classrooms. In: Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications. 2022.
87. Tadesse MM, Lin H, Xu B, Yang L. Detection of suicide ideation in social media forums using deep learning. Algorithms. 2019;13(1):7.
88. Thieme A, Belgrave D, Doherty G. Machine learning in mental health: A systematic review of the HCI literature to support the development of effective and implementable ML systems. ACM Trans Comput-Hum Interact. 2020;27(5):1-53.
89. Turecki G, Brent DA. Suicide and suicidal behaviour. Lancet. 2016;387(10024):1227-39.
90. Turecki G, Brent DA, Gunnell D, O'Connor RC, Oquendo MA, Pirkis J, *et al*. Suicide and suicide risk. Nat Rev Dis Primers. 2019;5(1):74.
91. Valeriano K, Condori-Larico A, Sulla-Torres J. Detection of suicidal intent in Spanish language social networks using machine learning. Int J Adv Comput Sci Appl. 2020;11(4).
92. Weidinger L, Mellor J, Rauh M, Griffin C, Uesato J, Huang PS, *et al*. Ethical and social risks of harm from language models. arXiv. 2021.
93. Wies B, Landers C, Ienca M. Digital mental health for young people: a scoping review of ethical promises and challenges. Front Digit Health. 2021;3.
94. Xu X. Detecting suicide ideation in the online environment: A survey of methods and challenges. IEEE Trans Comput Soc Syst. 2021;9(3):679-87.
95. Yang B, Lin X, Liu L, Nie W, Liu Q, Li X, *et al*. A suicide monitoring and crisis intervention strategy based on knowledge graph technology for "tree hole" microblog users in China. Front Psychol. 2021;12.
96. Yeskuatov E, Chua S, Foo L. Leveraging Reddit for suicidal ideation detection: a review of machine learning and natural language processing techniques. Int J Environ Res Public Health. 2022;19(16):10347.
97. Young T, Hazarika D, Poria S, Cambria E. Recent trends in deep learning based natural language processing. IEEE Comput Intell Mag. 2018;13(3):55-75.
98. Zhang T, Schoene A, Ananiadou S. Automatic identification of suicide notes with a transformer-based deep learning model. Internet Interv. 2021;25:100422.
99. Zhang T, Schoene A, Ji S, Ananiadou S. Natural language processing applied to mental illness detection: a narrative review. NPJ Digit Med. 2022;5(1).
100. Zhong QY, Mittal LP, Nathan MD, Brown KM, Knudson González D, Cai T, *et al*. Use of natural language processing in electronic medical records to identify pregnant women with suicidal behavior: towards a solution to the complex classification problem. Eur J Epidemiol. 2019;34(2):153-62.