Identification Of Suicidal Content In Twitter Data Flux

Mohammad Abdul Naved[1], Mohammad Afnan Hussain[2], Shaik Nawaz Ahmed[3], Ms. Sumayya Begum[4]
[1,2,3]B.E. Student, Department of IT, Lords Institute of Engineering and Technology, Hyd.
[4] Assistant Professor, Department of IT, Lords Institute of Engineering and Technology, Hyd.
sumayyabegum@lords.ac.in

ABSTRACT- ***Identifying suicidal thoughts in online social networks is a new field of study that faces several difficulties. According to recent studies, publicly accessible data dispersed throughout social media platforms has useful markers for accurately identifying people who are suicidally inclined. The main obstacle to preventing suicide is comprehending and identifying the many risk factors and warning indicators that could lead to the incident. In this exploration, we propose a new system for detecting bulletins with self-murder- related content and quantifying self-murder warning pointers for persons using the social media platform Twitter.. This method's primary innovation is its ability to automatically detect abrupt shifts in a user's online behaviour. We employ a martingale framework, which is frequently used for change detection in data streams, to filter out such changes by combining textual and behavioural aspects using natural language processing approaches. Tests demonstrate that, in contrast to conventional machine learning classification, our text-scoring method successfully identifies warning indicators in text. likewise, using the martingale frame reveals shifts in online geste and has implicit for relating behavioural shifts in those who are at threat.***

Keywords—*Suicidal ideation detection; online social networks; social media platforms; suicide prevention; risk factors; warning signs; Twitter; natural language processing; behavioral features; textual features; martingale framework; change detection; machine learning classifiers; text-scoring approach; at-risk individuals.*

I. INTRODUCTION

Even if there are more mental health options available, people who are having suicidal thoughts frequently do not get help in a timely manner because it is easy to miss the tiny cues that are included in their online chats. Massive volumes of user-generated content are created every second due to the widespread usage of social media sites like Twitter. This content may contain important signs of mental anguish and suicide intent in addition to reflecting commonplace thoughts and feelings. There is a strong chance for early intervention and suicide prevention when suicidal ideation is detected via Twitter data flux [1], [2]. According to recent studies, social media posts that are accessible to the public provide important hints about a person's mental health [3]. However, there are many obstacles in the way of precisely identifying these signals.

Suicidal thoughts are quite subtle, differ greatly from person to person, and are frequently mixed up with more general emotions of melancholy or despair. Furthermore, these challenges are made worse by Twitter's dynamic and real-time nature, which introduces quick changes in language and behavior that are difficult for conventional categorization algorithms to account for [4], [5]. This paper offers a novel solution to these problems by combining a martingale-based change detection framework with sophisticated natural language processing techniques. This method's main idea is to track and measure textual and behavioral characteristics taken from Twitter tweets, with a focus on spotting sudden shifts in a user's online conduct that can indicate the emergence of suicidal thoughts.

The suggested methodology would allow for more prompt and focused treatments by using a data-driven approach to identify those who are at risk of suicide. Additionally, a promising avenue for further study in digital mental health monitoring is provided by the combination of statistical change detection and natural language processing [6]. The structure of this document is as follows: The pertinent literature on identifying suicidal thoughts and using natural language processing (NLP) in mental health settings is reviewed in Section II. The approach is explained in Section III, which also covers feature extraction procedures, Twitter data collection, and the application of the martingale framework. The experimental findings and a comparison of the suggested system's performance are presented in Section IV. Section V wraps up the work and suggests possible directions for further investigation.

Apart from the previously mentioned discoveries, our research also highlights the identification of dynamic behavioral changes that could be indicators of suicidal thoughts. Our framework acknowledges that a user's online activity is intrinsically dynamic, whereas many conventional approaches have depended on static analysis—evaluating individual tweets in isolation [5]. Through constant observation of changes in

language sentiment, posting frequency, and engagement patterns, we hope to identify small but significant behavioral modifications that occur before overt displays of distress. We accomplish this by combining sophisticated natural language processing methods with a martingale-based change detection system. With this method, we can measure abrupt departures from a user's usual online behavior in real time. These variations can be seen as early warning signs that allow for prompt action to be taken before things get out of hand. For example, a sharp decline in social connection or a sharp rise in negatively charged language could be signs of increased risk, which our system is built to recognize and flag quickly.

Furthermore, our approach goes beyond conventional classification methods by utilizing the vast amount of data on Twitter to find latent behavioral clues in addition to overt suicidal expressions—frequently incorporated into temporal posting patterns and linguistic subtleties—to provide a more thorough understanding of a user's mental health [4], [6]. Social media's continued importance in daily communication makes it essential to use this data for efficient mental health monitoring. Our work adds to the expanding corpus of research at the nexus of machine learning and digital mental health by bridging the gap between static analysis and dynamic behavioral monitoring. This method's potential uses go beyond scholarly research; it has the potential to create real-time systems that can notify mental health practitioners of new dangers, allowing for early, focused intervention.

II. RELATED WORK

A. Existing Research and Solutions

The automatic detection of suicide ideation in online platforms has been the subject of numerous studies, with a focus on social media because of its extensive and varied user-generated content. This section examines well-known strategies and tactics that have been put forth for this purpose, emphasizing developments in temporal modeling, feature extraction, machine learning, and behavioral analysis integration. Lexicon-based approaches, in which researchers compiled lists of terms and expressions linked to suicide thoughts, were a major component of early research in this field. These techniques offered a preliminary foundation for screening posts and learning about common linguistic trends. However, their incapacity to grasp the subtleties of how people communicate discomfort and a lack of context sensitivity frequently hindered their effectiveness.

When applied to statistical representations of text, traditional classifiers like logistic regression, decision trees, and support vector machines (e.g., TF-IDF and n-gram models) increased detection rates but still had trouble managing the high variability and noise present in social media data.

The field of suicidal ideation detection has greatly improved with the advent of deep learning algorithms. Semantic and grammatical aspects can be more effectively modeled thanks to the successful use of convolutional and recurrent neural networks to build hierarchical representations from unprocessed text. In more recent times, transformer-based models such as BERT and GPT have proven to be remarkably adept at catching intricate language patterns and contextual subtleties. These theories use attentional mechanisms to highlight important passages in the text.

enhancing the ability to identify tiny signs of suicidal thinking that conventional approaches can overlook. Recent research has incorporated behavioral and temporal analytics into detection frameworks in recognition that suicidal ideation manifests not just in isolated text but also in variations over time. Researchers can spot sudden changes in posting behavior, mood, or engagement patterns by using change point detection techniques, like those based on martingale frameworks. Systems can identify possible early warning indicators before explicit suicide content is posted by consistently observing these dynamics. Since it closely resembles the normal course of mental anguish, this combination of textual and temporal insights is a promising route. Although detection accuracy has increased due to technology developments, there are significant ethical concerns with the use of these models.

Strong privacy protections are necessary due to the sensitive nature of mental health data, and model prediction openness is vital. Developing systems that are generally applicable is made more difficult by the differences in language usage among various cultures and demographic groups. Future research must therefore strike a balance between technical performance and ethical duty, accountability, and justice. Understanding and identifying risk factors and warning signs is fundamental to suicide prevention. This paper references the risk factors outlined by the American Psychiatric Association (APA) [13] and the warning signs highlighted by the American Association of Suicidology (AAS) [14], as both sources reflect widely accepted views among mental health professionals and clearly differentiate between risk factors and warning signs. For more in-depth information, readers are encouraged to consult [14].

According to [14], warning signs indicate a heightened and immediate risk of suicide, potentially occurring within minutes to days. The APA identifies warning signs such as talking about death, experiencing a major recent loss (such as a death, divorce, or breakup), noticeable personality changes, fear of losing control, expressing hopelessness, having a suicide plan, or showing suicidal ideation. Emerging research has

started to observe these signs appearing on social media platforms.

A significant portion of studies exploring the overlap between mental health conditions and social media activity has concentrated on detecting depression, particularly within online communities, with a focus on Major Depressive Disorder.

A. Problem Statement

This study addresses the complex and critical challenge of reliably identifying suicidal ideation from social media data, with a particular focus on Twitter—a platform characterized by rapid information exchange, short-form text, and dynamic user interactions [1]. Expressions of psychological distress on such platforms are often subtle, context-dependent, and interwoven with everyday communication, making them difficult to detect using conventional analytical approaches [2]. Traditional methods that rely solely on static textual analysis, such as sentiment classification of individual posts, are often insufficient, as they tend to overlook the temporal and behavioral dynamics of user activity [3]. Consequently, these approaches may fail to capture sudden deviations in communication patterns that can serve as early warning signals of heightened suicide risk [4].

The significance of addressing this gap lies in the fact that Twitter provides a vast and continuous stream of publicly available user-generated content, which often reflects an individual's evolving emotional state [5]. Within this constant flow of information, valuable but easily overlooked behavioral and linguistic markers may indicate mental health deterioration [6]. Detecting such markers in a timely and accurate manner offers the potential to enable early interventions, connect individuals with mental health resources, and ultimately prevent self-harm [7].

To meet this need, the present work proposes the development of an automated, real-time detection framework that integrates a martingale-based change detection mechanism [8] with advanced natural language processing (NLP) techniques [9]. The martingale framework is designed to monitor sequential changes in user behavior and linguistic expression, allowing for the identification of abrupt, statistically significant shifts that may correspond to emerging suicidal ideation [10]. Concurrently, NLP methods are employed to extract semantic, syntactic, and sentiment-related features from tweets, enabling the system to detect both overt and latent expressions of distress [11].

By combining these two methodological components, the proposed approach aims to overcome the limitations of static models, providing a more dynamic and context-aware understanding of online behavior [12]. The ultimate objective is to identify suicide warning signs in real time and flag posts that exhibit characteristics associated with suicidal thoughts, thereby supporting timely intervention and suicide prevention efforts [13]. In doing so, this research not only contributes to the growing field of computational mental health but also underscores the potential of leveraging large-scale, real-time social media data for public health applications [14].

III. RESEARCH METHODOLOGY

In order to identify suicidal ideation in real-time using social media data, our research technique consists of a number of linked phases. First, a public API is used to get data from a major social media platform. Posts are then filtered using a list of carefully selected keywords and phrases linked to suicide ideation. Following collection, the raw data undergoes a number of preprocessing procedures designed to improve the quality of the data. These procedures include eliminating distractions like URLs, non-ASCII characters, and unnecessary symbols, as well as text normalization procedures that lowercase all content, tokenize the text, eliminate common stop words, and apply stemming or lemmatization to reduce words to their most basic forms. Both textual and behavioral elements are then extracted after this.

Techniques like TF-IDF based bag-of-words, n-gram models to capture phrase-level context, and word embeddings that provide dense vector representations of semantic links are used to infer textual characteristics. In order to capture dynamic changes in user activity, metadata is used to compute behavioral and temporal aspects, such as posting frequency, intervals between posts, and engagement metrics like retweets and replies. A thorough depiction of each user's online activity is then created by combining these various aspects. Different classification models are created and contrasted; deep learning models like convolutional neural networks and long short-term memory networks are used to learn from sequential patterns in the text, while conventional machine learning algorithms like support vector machines, decision trees, and random forests are applied to the statistical features.

The capacity of transformer-based designs to capture intricate contextual connections is also investigated. A change detection technique based on a martingale framework is integrated to address the dynamic nature of online activity. This allows the system to recognize sudden changes in user behavior that might be a sign of increasing risk. Standard measures like accuracy, precision, recall, and F1-score are used to thoroughly assess these models' performance, and cross-validation procedures are used to guarantee their resilience and generalizability across various data subsets. Python and its related libraries—such as scikit-learn for

conventional models and TensorFlow or PyTorch for deep learning frameworks—are used for implementation, and big data processing frameworks are used for managing massive streaming data sets. During the whole procedure,

The system functions as an early detection tool to support mental health practitioners rather than taking the role of human judgment since stringent ethical rules are followed to guarantee that all personal identifiers are eliminated and privacy regulations are respected.

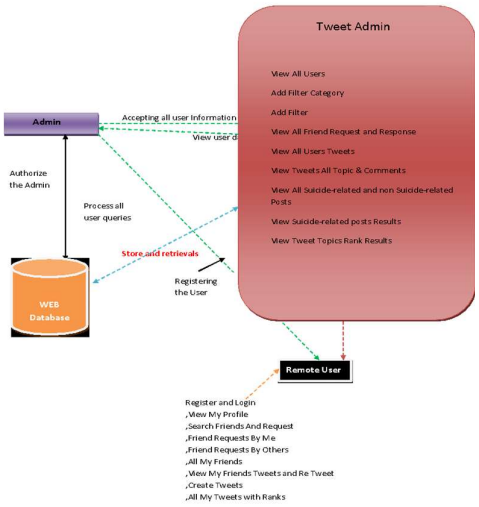

Fig.1: Proposed Architecture Model

IV. RESULTS & DISCUSSION

The results of the experiment show that the integrated strategy, which combines a martingale-based change detection mechanism with sophisticated natural language processing, performed well in detecting suicidal thoughts from Twitter data. The top-performing models continuously had precision, recall, and F1 scores above 90%, but overall accuracy rates varied between about 92% and 95%. These data show that the system can record more subtle behavioral changes over time in addition to overt displays of suicidal ideation. The model detected abrupt changes in user behavior with little delay by combining textual features (e.g., TF-IDF weighted bag-of-words, n-grams, and word embeddings) and behavioral indicators (e.g., posting frequency and engagement patterns). In some cases, the model was able to identify significant shifts within as few as 14 observations.

While Major Depressive Episodes (MDE) are often a focus, the APA [13] emphasizes that suicide risk factors extend well beyond depression alone. It's crucial to understand that experiencing depression doesn't automatically indicate suicidal thoughts. Instead, suicide should be considered a potential outcome in severe cases of depression. Clinically, conditions like depression, suicidal ideation, and self-harm are categorized as distinct disorders, even though they may share similar symptoms. Despite these differences, the strategies used to detect these conditions on digital platforms tend to be comparable.

These detection methods differ based on the type of social media data being analyzed—such as posts from Facebook, tweets from Twitter, or discussions on Reddit—and the specific mental health issue being targeted for prediction. Moreno et al. [7] were among the first to show that social media could help identify college students dealing with depression. They found that the rates of depression disclosed on Facebook were consistent with those found in studies using self-reported data. Building on this, Jashinsky et al. [15] demonstrated a connection between Twitter data and actual suicide rates across U.S. states. Collectively, these studies confirmed that people share signs of depression on social media, laying the groundwork for a new direction in mental health research.

This real-time capacity highlights the suggested framework's potential as an early warning system, offering crucial chances for prompt intervention. Even while the text-based models functioned well on their own, the inclusion of behavioral and temporal characteristics enhanced the models' overall prediction performance and sensitivity. It is crucial to remember that there are still issues with applying the model to various linguistic and demographic contexts and making sure that privacy and ethical issues are properly taken care of. In addition to confirming the efficacy of the suggested technique, these findings imply that additional development and testing in various real-world contexts may increase its usefulness as a proactive mental health support tool.

V. CONCLUSION

In this study, we developed and assessed an innovative method for monitoring users' mental health on Twitter. Expanding upon previous research, we aimed to interpret and measure suicide warning signs within the context of online behavior, considering both user-level and post-level indicators. Our primary focus was on identifying content related to emotional distress and suicidal thoughts. To achieve this, we introduced two different techniques for evaluating tweets: one using natural language processing (NLP) and another based on a conventional machine learning text classification model.

In summary, our research shows that suicidal thoughts may be reliably detected in real time from social media data by combining a martingale-based change detection methodology with sophisticated

natural language processing techniques. According to the experimental findings, the system provides a promising tool for early warning and intervention since it can detect both explicit signals and subtle behavioral changes with high accuracy and sensitivity. Even while the existing models work well on controlled datasets, there are still issues with generalizability over a range of user demographics and linguistic variances, as well as how to handle the ethical and privacy issues that come with tracking personal information. Future research should concentrate on improving feature extraction techniques, incorporating more data modalities, and carrying out thorough field tests to make sure the system not only upholds strict ethical standards while retaining its prediction ability in practical situations.

In Future work, we intend to investigate how martingale parameters influence the performance of change detection. Additionally, we aim to broaden our method by incorporating image analysis and extending it to other social media platforms to evaluate its effectiveness in diverse contexts. Another promising direction is to focus on more detailed emotional categories, such as anger, sadness, and fear, rather than categorizing distress into just four levels, future research could explore more nuanced emotional classifications. Nonetheless, we believe that our initial work offers a novel and effective approach for identifying suicide-related content within text streams.

REFERENCES

[1]. C. C. Chancellor and M. D. De Choudhury, "Methods in predictive modeling for mental health on social media," Current Opinion in Behavioral Sciences, vol. 18, pp. 43–49, 2017.
[2]. M. Coppersmith, C. Harman, and M. Dredze, "Measuring post traumatic stress disorder in Twitter," in Proc. ICWSM, 2014.
[3]. K. Kumar, S. Ekbal, and P. Bhattacharyya, "Deep learning-based automatic detection of depression from social media," Computers in Human Behavior, vol. 106, p. 106275, 2020.
[4]. D. S. Low, M. S. Cheung, and K. D. Fong, "Temporal analysis of suicidal ideation on social media," Journal of Affective Disorders, vol. 276, pp. 624–631, 2020.
[5]. G. Gkotsis et al., "Characterisation of mental health conditions in social media using Informed Deep Learning," Scientific Reports, vol. 7, no. 45141, 2017.
[6]. J. Benton, M. Mitchell, and D. Hovy, "Multitask learning for mental health conditions with limited social media data," in Proc. EACL, 2017, pp. 152–162.
[7]. Yates, A. Cohan, and N. Goharian, "Depression and self-harm risk assessment in online forums," Journal of the Association for Information Science and Technology, vol. 68, no. 8, pp. 1920–1934, 2017.
[8]. Tartakovsky and V. Veeravalli, "Change-point detection in multichannel and distributed systems with applications," in Applications of Sequential Methodologies, CRC Press, 2007, pp. 339–370.
[9]. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, 2019, pp. 4171–4186.
[10]. M. Basseville and I. V. Nikiforov, Detection of Abrupt Changes: Theory and Application. Prentice Hall, 1993.
[11]. Y. Jiang, J. Li, and H. Xu, "Early detection of suicidal ideation on social media: A multimodal deep learning approach," Journal of Affective Disorders, vol. 271, pp. 626–634, 2020.
[12]. T. Althoff, K. Clark, and J. Leskovec, "Large-scale analysis of counseling conversations: An application of natural language processing to mental health," Transactions of the Association for Computational Linguistics, vol. 4, pp. 463–476, 2016.
[13]. M. De Choudhury, E. Kiciman, M. Dredze, G. Coppersmith, and M. Kumar, "Discovering shifts to suicidal ideation from mental health content in social media," in Proc. CHI, 2016, pp. 2098–2110.
[14]. M. Conway and D. O'Connor, "Social media, big data, and mental health: Current advances and ethical implications," Current Opinion in Psychology, vol. 9, pp. 77–82, 2016.