JOIV : Int. J. Inform. Visualization, 7(4) - December 2023 2297-2305

INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv

Characteristics of Multiclass Suicide Risks Tweets through Feature Extraction and Machine Learning Techniques

Yan Qian Lim [a], Yim Ling Loo [b,*]

[a] Department of Internet Engineering and Computer Science, Universiti Tunku Abdul Rahman, Selangor, 43000, Malaysia
[b] Faculty of Computing and Informatics, Universiti Multimedia, Cyberjaya, Selangor, 63100, Malaysia

Corresponding author: *looyimling@mmu.edu.my

***Abstract*— This paper presents a detailed analysis of the linguistic characteristics connected to specific levels of suicide risks, providing insight into the impact of the feature extraction techniques on the effectiveness of the predictive models of suicide ideation. Prevalent initiatives of research works had been observed in the detection of suicide ideation from social media posts through feature extraction and machine learning techniques but scarcely on the multiclass classification of suicide risks and analysis of linguistic characteristics' impact on predictability. To address this issue, this paper proposes the implementation of a machine learning framework that is capable of analyzing multiclass classification of suicide risks from social media posts with extended analysis of linguistic characteristics that contribute to suicide risk detection. A total of 552 samples of a supervised dataset of Twitter posts were manually annotated for suicide risk modeling. Feature extraction was done through a combination of feature extraction techniques of term frequency-inverse document frequency (TF-IDF), Part-of-Speech (PoS) tagging, and valence-aware dictionary for sentiment reasoning (VADER). Data training and modeling were conducted through the Random Forest technique. Testing of 138 samples with scenarios of detections in real-time data for the performance evaluation yielded 86.23% accuracy, 86.71% precision, and 86.23% recall, an improved result with a combination of feature extraction techniques rather than data modeling techniques. An extended analysis of linguistic characteristics showed that a sentence's context is the main contributor to suicide risk classification accuracy, while grammatical tags and strong conclusive terms were not.**

***Keywords*—Multiclass suicide risks; suicide ideation detection; feature extraction; machine learning; sentiment analysis.**

Manuscript received 5 Dec. 2022; revised 29 Jul. 2023; accepted 25 Aug. 2023. Date of publication 31 Dec. 2023.

I. Introduction

An individual exchange of information is in the form of expressing thoughts, feelings, or desires through verbal or non-verbal elements, writings, sketches, or paintings in virtual or physical communication mediums. A radical shift was observed in present individuals, utilizing virtual communication mediums such as Facebook, Reddit, Twitter, Instagram, and TikTok as the primary medium of information exchange and acquaintance establishments [1]– [5]. Nonetheless, organizations and institutions of different domains were observed using similar virtual communication mediums and social media platforms to publish information, promotions, and advertisements [6]– [9]. Thus, such social media platforms are an essential medium of communication for individuals and organizations.

Familiarity with using social media platforms for communications in different areas of life motivated individuals to consider it a haven to convey honest thoughts without being criticized or judged freely. As such, a significant growth of individuals who would choose to express their feelings and ideas, especially suicide impulses on social media platforms, was observed [1]–[3], [10]–[12]. Consecutively, social media platforms turned into a handy and popularized suicidal content-sharing point. [5] reported that youngsters were found to be more earnest in communicating suicidal thoughts through social media platforms, for instance, Facebook and Twitter, yet being apathetic during medical appointments with mental health professionals.

Various research studies have been undertaken in which different Artificial Intelligence approaches have been explored to understand its role and efficacy in detecting suicide ideation among social media posts. However, these studies are mostly limited to exploring and reporting the feasibility and optimization of efficiency through the implementation of Artificial Intelligence techniques such as

deep learning, machine learning, and feature extraction approaches [3], [4], [13]–[18]. Such studies were observed to have further focused on expounding possibilities for improving binary classification of suicide ideation detection.

These studies delivered either deep learning or machine learning models that detect suicide ideation on textual posts, where they further classified the detected suicide ideation as either suicidal or non-suicidal context only. Moreover, these studies focused on integrating and comparing implementations of black-box algorithms to improve model's performance without scrutinizing the underlying factors or features within the textual social media posts that impacted the performance of the models. In other words, there is a lack of studies that explore the patterns associated with different degrees of suicidality in correlation to its impact on the suicide ideation detection model's performance [18], [19].

With prior research studies verifying the efficacy of taking an Artificial Intelligence approach towards this problem domain, there is reason to leverage this opportunity to explore at a lower granularity how patterns associated with different levels of suicide risks affect the efficiency and accuracy of the model in discerning different levels of suicide risks. Likewise, different aspects constitute the syntactic structure of suicidal speech. Hence, understanding these characteristics will contribute to knowing which algorithm or feature should be prioritized for more effective and efficient suicide ideation detection. Binary classifications of suicide ideation and comparison between black box approaches alone are not sufficient for such a proposed scope of studies.

In this context, we proposed a machine learning framework [20], which performs multi-classification of suicide risks from detected suicide ideations in Twitter posts in the endeavor of a better suicide ideation prognosis from textual social media posts. This paper documents the implementation of the proposed framework not only to prove the reliability of results produced by the proposed approach but also to leverage correlations between syntactic patterns found in suicide ideations and suicide risk levels.

In particular, this work aims to address research questions. *RQ1:* How various linguistic characteristics can be extracted from suicidal social media posts? *RQ2*: Which linguistic characteristics impact the most on model's prediction of suicide risk levels? To the best of our knowledge, this is the first attempt to implement various feature extraction techniques for linguistic characteristics extraction from suicidal social media posts dataset with extended multi-classifications of suicide risk levels to facilitate further studies on this domain. To facilitate this study, we implemented our proposed machine learning framework, proving the reliability and robustness of the data model through accuracy, precision, and recall results found on par with other studies. We analyzed the impact of various linguistic characteristics on classifications of suicide risk levels for findings of the most prominent and insignificant characteristics. The outcome of this analysis will further contribute to the efficiency and accuracy of suicide ideation detection in social media posts.

Previous studies have explored the implementation of an Artificial Intelligence approach for detecting suicide from social media posts. To that end, traditional machine learning models and deep learning methods were prevalently used for suicide classification on supervised learning datasets. [17] developed a hybrid model that combines the strength of artificial neural network methods to improve the classification performance. [14] explored the use of machine learning and ensemble approaches for classifying Twitter posts. [11] created a framework that applied feature extraction and machine learning to classify Reddit and Twitter posts. However, the studies presented above tested their approach for binary classification only, whereby the post was classified as either suicidal or non-suicidal. Consequently, it is inadequate as suicide ideation is an umbrella term that encompasses circumstances from wishing for one's death to making plans to commit suicide [21]. Hence, there are limitations to the discoveries from these studies as it does not cater to understanding passive and active suicide ideations.

In studies adopting multiclass classification, [22] presented an approach for classifying social media posts into 4 risk levels: no risk, low risk, moderate risk, and severe risk. The dataset achieved high annotation quality as it had low inter-rater disagreement between human assessors. Using Twitter posts, studies by [23] proved the feasibility of identifying different levels of suicidality from posts by using machine learning techniques to replicate the accuracy of human coders. The authors defined a three-level categorization criterion and discovered that most of the tweets within their collected dataset contained some level of concern. However, these studies did not explore the pivotal characteristics that are associated with each type of suicide risk. These characteristics are crucial as they help the model learn from these associations to form determination points for predicting future suicidality. As a result, this study addresses the existing research gap by exploring the characteristics and patterns significantly associated with different levels of suicide risk and demonstrating their contribution to the model's predictive ability.

Feature extraction is a crucial approach in machine learning because it extracts key information from a text input and converts it into a feature set in the classifier [24]. For textual data, natural language processing techniques are applied to analyze and process the unstructured data to identify the key data attributes that provide high-quality information. However, the specific techniques used impact the feature's quality, significantly affecting the classification outcome [2]. Studies by [23] demonstrated that applying the Term Frequency-Inverse Document Frequency (TF-IDF) allowed the model to perform better when compared to other feature extraction techniques, such as simple frequency. [11] used several types of feature extraction techniques, including linguistic inquiry and word count, TF-IDF, and part-of-speech (PoS) tags, to extract features from a Reddit dataset. [22] also used data acquired from Reddit and applied Bag of Words (BoW) to convert the data into a vector representation and TF-IDF to assign the word weights. However, because TF-IDF balances the frequency of common and uncommon terms, it is more extensive than BoW. In comparison, BoW only computes the frequency of terms in the given text, resulting in domain-specific words that hold greater significance but are less frequently used to be overlooked [25]. Therefore, the proposed feature extraction approach will use a combination of TF-IDF and PoS tagging to extract the features from the Twitter dataset.

Sentiment analysis, which seeks to understand the emotional perception underlying a particular text, was also a common feature used in research studies. [26] used Linguistic Inquiry Word Count matrices and sentiment matrices to form their feature sets in this context. In another study, [23] extracted sentiment features by computing the percentage of sentences or terms associated with a specific polarity. This study will use sentiment analysis to understand the sentiment expressed within the tweet.

Random Forest and Support Vector Machine (SVM) were the most commonly used machine learning models used in previous works. In works by [14], it was found that Random Forest had the best performance among other modes, including Naïve Bayes, SVM, and Logistic Regression (LR). According to studies by [16], which used a large feature set as an input for the model, Random Forest also achieved the best performance results among other models, such as Bayesian Network, Adaboost, and J48. This is as Random Forest adopts a dimensionality reduction method to filter and identify the most significant variables in the dataset, thus reducing the risk of overfitting the model in the presence of large datasets with large feature vectors [27]. A study by [15] found SVM achieved the best performance among other models, including Random Forest and LR. Similarly, [23] found that SVM performed the best in their study. However, the model could not achieve a learning plateau when more data is added, indicating that a larger dataset is needed for the model to achieve its peak performance. Given the nature of the training data, which is noisy and contains a high number of features, Random Forest was utilized in this study, considering its proven ability to perform well on similar datasets with these characteristics.

Based on the research studies presented above, it is concluded that a large number of studies leverage different techniques to explore this problem domain from multiple perspectives. Despite their unstructured and arbitrary nature, social media posts follow regular language patterns that can be useful learning information for the model. To that end, this study provides further exploration and extension to existing research by identifying the characteristics that exist within the boundaries of each suicide risk and demonstrating its impact on the model performance. Leveraging validated feature extraction techniques from previous studies, this study outlines how these characteristics coincide to form patterns that are beyond human intuition, which helps the model be more effective in its classification task. The approach was tested on supervised Twitter datasets to obtain more significant insights into the model's performance [28].

II. Materials and Method

This section outlines the implementation of a formulated machine learning framework for suicide ideation detection in Twitter, a preliminary finding of this study as expounded in [20]. Thus, details of methods that were already reported previously will not be repeated in this paper.

A. Data Collection

This study used a manual approach to build the dataset because public datasets on suicide-indicative tweets are limited due to privacy concerns, as declared in the courtesy of data usage on the Twitter platform before data collection permission approval. Previous studies have used Twitter to collect tweets to detect hate speech and analyze COVID-19 vaccine discussions and mental health characteristics [29]–[32]. The tweets were collected through Twitter API using a keyword filtering approach, in which a list of suicide-indicative keywords was passed into the search query to capture tweets that matched such keywords. The keyword filter was based on terms validated by more than 80% of the respondents in the works of [33].

Therefore, a total of 22 keywords were used for the filter, which comprised of terms or phrases such as "better off dead," "slit my wrists," "suicide," and "suicidal ideation". The keywords were further refined as it was found that some terms included in the query, such as "fleeting thoughts of suicide" and "completed suicide," were less conventional among online communication and were thus ineffective in yielding relevant results. Therefore, the final list of keywords contains a total of 16 keywords, which are: "better off dead," "blow my brains out," "blow my head off," "commit suicide," "contemplating suicide," "hang myself," "kill myself," "self-harm," "shoot myself," "sleep forever," "slit my wrists," "suicide," "suicidal," "suicidal ideation," "suicidal thoughts" and "want to die." 4,000 tweets were collected from June 9, 2022, to June 19, 2022, including attributes such as full tweet text, username, date and time of a tweet, location name, and location coordinates. However, usernames are concealed for ethical considerations, while geolocation information, i.e., location name and coordinates, are not in the scope of this study. Geolocation information is best to be considered for the use of extensive study of this research.

B. Data Annotation

The collected dataset was annotated based on the adaptation from [22] and [23], where risk criteria classified (0) Low Risk: No evidence or patterns that implicate the user is at risk of suicide; (1) Medium Risk: Possible suicide risk is identified from the user content, but no emergency assistance required; (2) High Risk: Strong and decisive phrases which implicate serious suicidal intent, where emergency assistance is urgently required. Initial analysis showed that the majority of the collected tweets were found to be of low risk as they primarily contained references to either television series, movies, or news reports. Hence, these low-risk tweets were excluded from the dataset to reduce model bias due to imbalanced class distribution. Duplicate data is removed to avoid distorted outputs. As such, the final annotated dataset contains 690 tweets, with an even distribution of 230 tweets for each risk level. Table I shows examples of tweets from each suicide risk level.

TABLE I
TWEET SAMPLES OF EACH SUICIDE RISK

Risk Level	Tweet
(0): Low Risk	• The weather is nice and warm today, might even shoot some pictures by myself too • I am just happy to hang out with new cool and fun people instead of being by myself
(1): Medium Risk	• Really wish I could shoot myself sometimes • Caught myself thinking that I am better off dead
(2): High Risk	• God, I want to kill myself so bad right now • I will blow my brains out

C. Data Pre-Processing

The annotated dataset was pre-processed to remove redundant characters, including usernames, punctuations, hyperlinks, special symbols, numbers, and additional whitespace. Although studies by [10] and [22] consider emojis and emoticons in their cleaned dataset, these were removed in the present dataset due to their inconsistent data format. Additionally, each contraction was resolved to its original group of words to standardize its expression across the dataset. Then, tokenization and lemmatization were carried out to filter and normalize the words through Python's Natural Language Toolkit (NLTK). An example of a pre-processed tweet is shown in Figure 1.

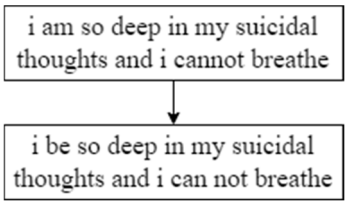

Fig. 1 Pre-processed Tweet

D. Feature Extraction

Feature extraction was executed to capture the significant features for constructing the final feature set, which will serve as input for the model [3], [10], [20], [34].

1) Term Frequency-Inverse Document Frequency (TF-IDF): TF-IDF assigns a low score to irrelevant terms while assigning a higher value to terms that convey crucial semantic information. This reduces the classification model's training time because the model may prioritize words that contain crucial information within the tweets. For this study, the TF-IDF vectorizer was configured to adopt a unigram approach, with an additional configuration to omit exceptionally uncommon terms that are found in less than five tweets. Removing stop words might contribute to better TF-IDF scores. However, in suicide detection, works by [35] found that additional pre-processing might negatively impact model performance, as it contributes to the distortion of critical information due to drastic changes in the textual context.

2) Part-of-Speech (PoS) Tagging: PoS tags were used as features to help the model identify patterns associated with the grammatical characteristics of the tweets. Each tokenized tweet was evaluated against 35 PoS subgroups and labeled with its respective tag, defined internally by the NLTK library. The PoS tags provide information about the corresponding grammatical subgroup of each word. The total number of words in each grammatical category was then calculated.

3) Sentiment Analysis: Sentiment analysis was performed using NLTK's Valence Aware Dictionary and Sentiment Reasoner (VADER) library to analyze the emotional sentiment associated with each tweet. VADER's sentiment analyzer function generates four semantic scores by default: positive, neutral, negative, and compound. The compound score is the normalized total of the negative, neutral, and positive scores on a scale that ranges between -1 and 1. Hence, the compound score is passed as a feature in the dataset.

E. Model Training and Testing

An 80:20 split ratio on the dataset was applied, whereby 552 samples were used for training, and the remaining 138 samples were used for testing. To mitigate the risk of model bias, the risk label was stratified to include an equal percentage of each class size in both the training and testing sets. The training set is then fed into the Random Forest classification model, which is validated against the hold-out test set.

This work acknowledges that for the problem area of suicide monitoring, accuracy alone is insufficient to evaluate model performance because it only analyses the model's ability to categorize risk levels accurately. Other factor, such as recall, measures the model's ability to identify positive suicidal risks should be prioritized to ensure that the model does not overlook suicidal content [26]. Precision must also be considered to ensure that the model can effectively rule out non-suicidal content. Therefore, a combination of the three-performance metrics, accuracy, precision, and recall, was used to assess the model performance and ensure that the model maintained an acceptable balance between these three metrics. The model performance evaluation is presented in the following section.

III. Results and Discussion

The results were analyzed based on two aspects. Firstly, a model performance evaluation was performed to understand the model performance on the hold-out set. The trained model was then evaluated based on random tweets that were obtained in real-time. The model's performance was evaluated using three main metrics: accuracy, precision, and recall. From these results, the patterns and findings outlined during data analysis were leveraged to understand their impact on the model's performance. It is worth noting that the model performance results are based on three extracted features. Following model performance results, an exhaustive analysis of the significance of linguistic characteristics contributing to the model's performance is reported. Then, the section is enveloped with the conclusive findings in addressing the research questions of this work.

A. Model Performance Evaluation

It was observed that the model achieved good performance results on the test set, achieving an accuracy, precision, and recall of 86.23%, 86.71%, and 86.23%, respectively, which is on par with the works of [13], [16], [17], [35], [36] as tabulated in Table II. The existing works listed in Table II investigated a dataset of texts exchanged in social media as well as classifying suicide risks into multiple classes, which are similar to this study. This ensures that the comparison is an accurate benchmark for this work. Figure 2 shows the confusion matrix of the model to gain further insight into the model's performance.

It was significant that medium suicide risk had the highest true positive rates, with two samples more than for low and high risk. False positives, on the other hand, were significantly higher across low and high suicide risk levels, as the model tends to predict low suicide risk to have medium to high levels of suicide risk and tweets with high suicide risk to have low and medium levels of suicide risk. As for false negatives, it was observed that there was a higher

misclassification among tweets with low and medium suicide risk.

TABLE II
MODEL PERFORMANCE COMPARISON WITH EXISTING WORKS

Prediction Models	Accuracy (%)	Precision (%)	Recall (%)
Model performance (this study)	86.2	86.71	86.23
Real-time performance (this study)	89.33	92.98	83.08
[35]	71 - 76	69 - 74	65 - 74
[13]	61.2 - 64.2	60.6- 62.7	62.5 -70.1
[16]	-	81	90
[17]	77.2 - 85.6	76.3 - 85	75.1 - 84
[36]	70	81	-

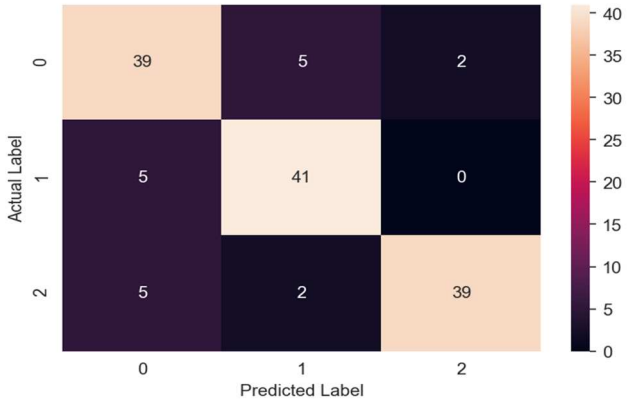

Fig. 2 Confusion Matrix of Model

The results showed that the feature set used to train the Random Forest model was effective, as the model could differentiate tweets across its risk levels by leveraging the pattern from the sentiment score and discussions in Section *III,* part *C*. However, from the reverse perspective, as there is no definitive threshold for the sentiment score of each risk level, the outliers would compromise the model's ability to predict the risk levels correctly, resulting in misclassification.

Although it was initially assumed that higher misclassification would occur within tweets of medium suicide risk due to the high variability of its sentiment score, the results showed that low suicide risk had the highest false positives, which is quite the contrary. This is because our implemented approach uses multiple feature extraction techniques to create the feature set. As such, the model also factors in those features when predicting the risk label of the tweets. In comparison, if only one feature is used, the model would be ineffective because of the high variability of its sentiment score since it is limited to that information for prediction. This demonstrates that using different feature extraction techniques to build the feature set contributes to the model's ability to perform its classification task effectively, consistent with the findings by [17]. A detailed analysis of the impact of TF-IDF and PoS tags is presented in the next section.

B. Real-Time Performance Evaluation

In order to obtain an impartial validation and evaluation of the model's performance on undiscovered data samples, the model was evaluated based on tweets that were captured in real-time from Twitter API. Unlike the original dataset used for model training, the real-time tweets went directly to phases of data pre-processing, feature extraction, and model prediction without prior intervention to remove completely unrelated samples. Hence, these samples mirror user input that is random and unfiltered, as the topic discussed in the tweets was arbitrary, consisting of thoughts, conversations, or news articles. The real-time samples were then manually annotated using the same risk categorization guideline defined previously to identify the discrepancy between the actual and predicted risk labels.

From there, it was found that the model could achieve high-performance results with accuracy, recall, and precision of 89.33%, 92.98%, and 83.08%, respectively, as tabulated in Table II. These findings show that the feature set used during classification, specifically the TF-IDF and sentiment score, affects the model's performance. Section *III* part *C* expounds on the significance of each linguistic characteristic through analyses of individual features of PoS Tags, TF-IDF, and sentiment scores for clear visualizations of the impact of feature set on classifications of suicide risk levels.

C. Linguistic Characteristics Evaluation

Sentiment analysis was carried out on each feature through various visualization techniques to explore and obtain foundational ideas on the underlying patterns of the dataset. The features include PoS Tags, TF-IDF, and Sentiment, which are expounded in the following sections.

1) PoS Tags: As mentioned in the previous section, PoS tags allow us to understand the grammatical characteristics of a given text. This context outlines how the grammatical style varies across different suicidal risks. To illustrate, Figures 3 and 4 demonstrate the ten most tagged PoS labels within tweets that are classified as medium and high suicide risk.

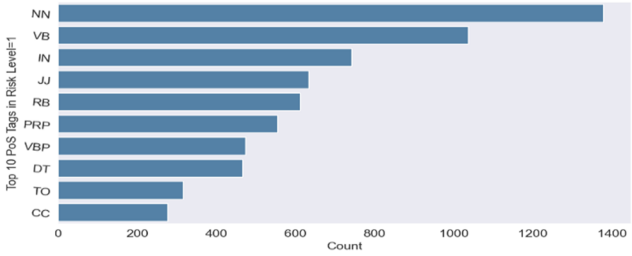

Fig. 3 Top 10 PoS tags of Medium Risk Tweets. NN is for Noun, VB is for Verb, IN is for Preposition, JJ is for Adjective, RB is for Adverb, PRP is for Personal Pronoun, VBP is for singular verb, DT is Determiner, TO is for To Go, CC is for Coordinating Conjunction

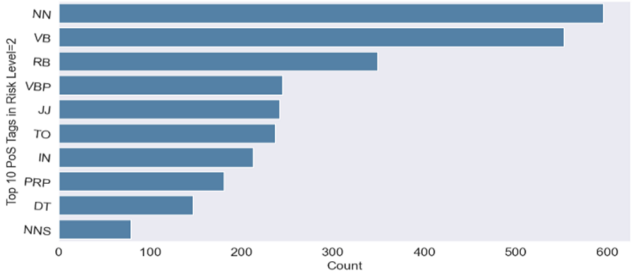

Fig. 4 Top 10 PoS tags of High Risk Tweets. NN is for Noun, VB is for Verb, RB is for Adverb, VBP is for singular verb, JJ is for Adjective, TO is for To Go, IN is for Preposition, PRP is for Personal Pronoun, DT is for Determiner, NNS is for Plural Noun

It is observed while comparing Figures 3 and 4, that a slight difference in syntactic features in terms of frequency and grammatical properties used was found in association with tweets of medium and high suicide risk. While tags in medium and high-risk tweets overlap, with a singular noun (NN) and base form verb (VB) being the most common, it is observed that there is a significant difference in the total count, which comprised the count for NN and VB in high risk.

2) TF-IDF: To obtain further insight into the TF-IDF word matrix in this study, the mean TF-IDF score is calculated to identify how the most significant terms differ across medium and high-risk levels. The top 25 terms of each risk level were then visualized in bar charts, which are illustrated in Figure 5 and Figure 6.

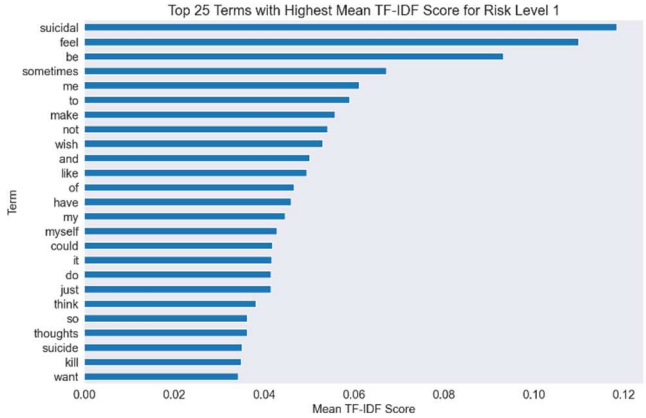

Fig. 5 Top 25 Terms with Highest Mean TF-IDF Score of Medium Risk Tweets

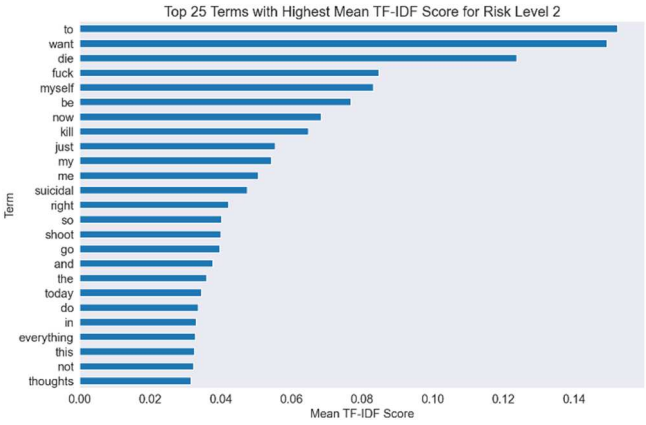

Fig. 6 Top 25 Terms with Highest Mean TF-IDF Score of High Risk Tweets

Although the terms used in both risk classifications overlap, they differ in terms like "suicidal", "feel" and "thought," which were prevalent in medium-risk tweets, whereas solid and determinative terms like "die," "kill," and "want" were dominant in high-risk tweets. This finding revealed that the terms that should be considered crucial for each suicide risk label observe a distinguished pattern, with suicide ideation-instigated terms commonly used among medium-risk tweets, while terms that are more proactive display an active determination to commit suicide being used in high-risk tweets. Additionally, the TF-IDF score disclosed that each suicide risk level's dominant terms were either nouns or verbs, which aligned with the findings highlighted in the previous sub-section.

3) Sentiment Features: Figure 7 shows the relationship between the sentiment score and each risk level.

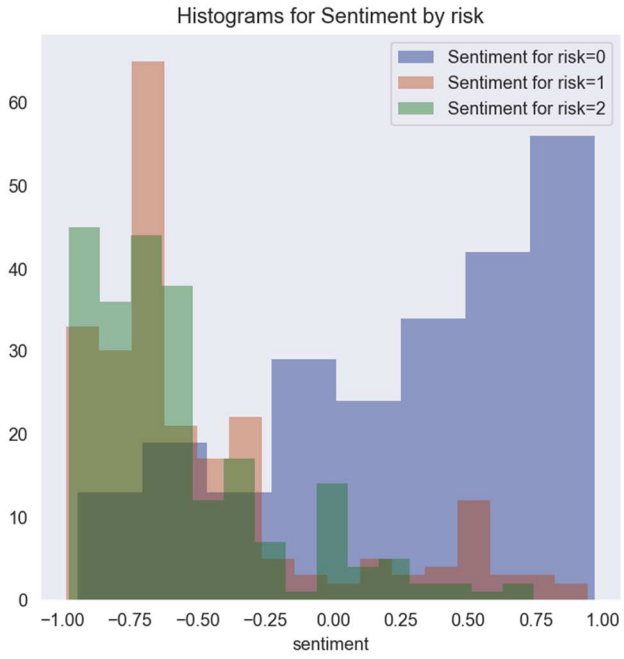

Fig. 7 Histograms of Sentiment Score by Suicide Risk Level

According to Figure 7, it is identified that the sentiment score for each suicide risk level is distributed across the spectrum. As a result, no distinct range can be defined for each risk level. However, it is observed that the samples follow a trend, where low suicide risk samples skewed towards sentiment scores that exceed -0.75 and high suicide risk samples skewed towards sentiment scores below 0.75. On the other hand, the sentiment scores for medium suicide risk were less consistent, as the values have random fluctuations across the scale. This could potentially impact the model's ability to identify the underlying connections between the feature and its target suicide risk during model training, thus resulting in misclassification.

Furthermore, it was found that the choice of words used within the tweet impacted the sentiment score from VADER. For instance, Figure 8 shows a tweet of medium risk label and its corresponding sentiment score.

```
i have a few not so bad days i think how long will this take until i slit my wrists
###VADER Score Listing###
neg 0.0
neu 0.841
pos 0.159
compound 0.5549
```

Fig. 8 Sentiment Scores of Medium Risk Tweet Sample

Despite conveying suicidal intent, a combination of neither excessively optimistic nor pessimistic terms resulted in a compound score that skewed towards the positive end of the scale. Therefore, sentiment intensity computed by VADER is found to be significantly dependent on the choice of words used in the tweet. For this reason, using sentiment score alone is insufficient for the model to make accurate predictions on the suicide risk. This finding highlights the importance of combining machine learning with feature extraction to determine suicide risks, which is further presented in the next section.

4) Extended Analysis for Real-Time Tests: Findings from some of the real-time testing datasets shaded lights to understanding the impacts of all three features that had been discussed in previous sub-sections on prediction results. Table III shows excerpts of correctly classified tweet samples of medium and high suicide risk.

TABLE III
EXCERPTS OF CORRECTLY CLASSIFIED TWEETS

Risk Level	Tweet
(0): Low Risk	• I recovered so much from suicidal ideation that I am at a point where I not just stopped wanting to kill myself but actively enjoy being alive. None of that "I wish I never was born". I am so glad I was born and got to live. It can get better, keep going • It's a mental illness, but considering we push people to live out this ill-ness, these people are really fragile and suicidal, so best we be very careful with our words... because that "what doesn't kill you makes you strong stuff" hurts people's feelings.
(1): Medium Risk	• Sometimes I think If?? I'm over the suicidal thoughts but then I could be doing the simplest of things or nothing at all and they come, and frequently now • It's been a while since I felt depressed or suicidal, I was so happy. Now everything is falling apart.
(2): High Risk	• Yep. Cut my wrist. I called my own ambulance. Was not a good time to be me. • So close to the edge. I do not know how much strength I have to bring myself up. but it doesn't matter because they all let me hang off for so long, I might as well blow my fucking brains off

It is observed that these tweets contain overlapping terms with terms that were considered substantial for its suicide risk level which was shown previously in Figure 5 and Figure 6. When checked against its sentiment score, its value was consistent with the pattern as shown in Figure 7. This therefore exemplifies the trait of the feature set for it enables the model to effectively differentiate between contexts of various suicide risk by leveraging the patterns revealed in the TF-IDF and sentiment scores.

Other than that, the model was found to be able to correctly classify tweets of low suicide risk, which included tweets that recalled previous suicide attempts, made indirect references to suicide and mental health, and used suicide indicative terms in non-suicide indicative contexts. According to the low-risk samples in Table III, it is observed that a compound of suicide-related keywords such as "suicidal ideation", "kill myself" and "suicidal" were used within the tweets. If TF-IDF is the only feature passed as an input to the model, there is a high possibility that misclassification would occur for such samples, as these keywords were identified to be strongly associated with medium- high suicide risk, although the whole context itself does not contain any suicide risk. However, the sentiment score computed for these tweets were inclined to positive end of the scale due to inclusion of positive connotated terms such as "enjoy", "glad" and "strong". This indicates that the use of multiple features allows the model to provide an accurate prediction of its suicide risk as it analyses the tweet with consideration towards its overall context. Hence, it is conclusively proven that the model effectively classifies the tweets due to its ability to analyze the patterns found among each individual feature set and balance its correlations to specific risk categories, thus allowing it to make accurate predictions.

However, it was found that the PoS tags alone do not have any significant impact on the model's performance. Table IV further illustrates the PoS tag scores of the correctly classified tweets from Table IV, which is truncated for better emphasis on the top 10 PoS tags identified during data analysis.

TABLE IV
POS TAG SCORES OF TWEETS

Risk Level	Sample No.	PoS Tags										
		CC	DT	IN	JJ	NN	NNS	PRP	RB	TO	VB	VBP
(0): Low Risk	i	2	2	4	5	13	0	2	7	2	10	4
	ii	2	4	2	5	4	3	5	4	1	5	5
(1): Medium Risk	i	4	3	3	3	2	2	1	4	0	3	2
	ii	1	1	1	2	6	0	1	3	0	4	0
(3): High Risk	i	0	1	0	2	6	0	1	1	1	2	1
	ii	1	2	4	2	7	0	4	7	2	5	3

It is observed that there were no significant differences between the PoS tags count of medium and high-risk levels. For instance, prepositions (IN) and adjectives (JJ) are either equally or more often used in high-risk tweets compared to medium risk tweets. Besides that, it is worthy to note that the count for singular noun (NN) and verbs (VB) was found to be highest for the first sample of low suicide risk, compared to the other samples since it was lengthier and used a variety of words within the context. Despite that, the higher count in singular noun and verbs does not directly indicate that the suicide risk level of the tweet is to contain medium or high risk. In other words, the grammatical tags do not have a strong association with specific suicide risks. As a result, PoS tagging alone has a low predictive power for the model to make inferences on the suicide risk level.

Besides that, it is also found that more context is needed in order for the prediction to be effective. To illustrate, Table V shows an excerpt of tweets that are misclassified by the model.

TABLE V
EXCERPTS OF MISCLASSIFIED TWEETS

No.	Tweet	Predicted Risk	Actual Risk
i	that's just some shit i cannot do	1	0
ii	fuck my life	1	0
iii	im so tired	2	0
iv	this is so fucking stupid lmao	2	0

According to Table V, the tweet samples conveyed negative general context, as these tweets use profanities and terms with negative connotation. Considering the strong language used as shown in Table V, these tweets would be given high negative scores as computed by VADER.

Moreover, there are no neutral or positive connotations to balance the context of the tweet sample as the length of the tweet is short. As a result, the tweets are bound to have a low compound score from VADER.

Furthermore, lack of context would also influence TF-IDF scores. Evaluation of tweet samples based on the terms that are found within the matrix is constrained, as TF-IDF features sets are established on word matrix that the vectorizer was pre-trained with. Furthermore, terms from these samples that overlapped with existing terms within the word matrix were strongly associated with negative emotions. As a result of the negative terms' impact on the TF-IDF score and VADER sentiment score, the model misclassifies the samples in Table IV as having a medium or high suicide risk.

D. Research Contributions

Based on the model performance evaluations and linguistic characteristics analyses, the two-fold research questions posed in introduction are ready to be answered.

1) [RQ1.], How various linguistic characteristics can be extracted from suicidal social media posts?

We have presented a three-fold feature extractions technique, which is an implementation machine learning framework for suicide ideation detection in Twitter which was formulated in preliminary works of this study. The three-fold features are namely PoS Tags (Syntactic feature), TF-IDF (word frequency) and sentiment feature with multi-classification of low, medium and high suicide risk levels. Promising results of model performance demonstrated the magnitude of the features in impacting the model's performance. Extensive analysis on the multi-features with multi-classifications of suicide risks contributed to a thorough analysis upon predictions of not only whether the Twitter post contain suicide ideation, but level of suicide risk when a suicide ideation is detected.

2) [RQ2.], Which linguistic characteristics impact the most on model's prediction of suicide risk levels?

We have presented the analyses of compound linguistic characteristics impact on model's prediction of suicide risk levels through testing of model with supervised testing datasets. We will begin answering this research question through extensive breakdown analysis on individual features to shed light on individual feature's dominance on model's performance. The noun and verb count of PoS tags show distinctive values which differentiate high risk from medium risk. Frequency of proactive verbs under TF-IDF significantly impacts predictions to be skewed to high risk while common suicidal terminologies to be medium risk. Low, medium and high suicide risks were found to be heavily on positive sentiment values, -0.75 to 0.5 sentiment values and negative sentiment values, respectively. However, it was found through analyses of real-time testing datasets predictions, three of the features could not impact the prediction as standalone features. The features need to be consolidated for better suicide risk level prediction. It was observed that the instances of lack of context which forced the model to fully depend on only one feature for analysis to prediction, the prediction would end up to be false positive or false negative. Thus, the more context the input to be fed into the model as how most of the testing datasets were, the model could predict better, as all the linguistic characteristics had compound analysis.

We presented an effective approach to train the classification model as it achieved high performance results that were consistent with existing works. Upon further analysis, positive associations were found between the patterns within each feature that relates to a specific suicide risk level. The model was able to accurately classify tweets by leveraging the pattern found in the VADER sentiment scores. This encompasses high accuracy of classifying low-risk tweets which contain indirect references to suicide-related keywords and recall previous suicidal experiences. At the same time, results show that the model was able to associate the patterns found within the TF-IDF score to discern key terms that are used to communicate different degrees of suicide ideation.

IV. CONCLUSION

Through this work, three feature extraction techniques were implemented: TF-IDF, PoS Tags and sentiment analysis to form feature set which is subsequently passed into Random Forest model for classification task, which yielded high performance results that were comparative with existing works. When tested with tweet samples captured in real-time, the model showed promising results, hence proving its ability to classify and predict undiscovered data and thus verifies efficiency of the model in real-time domain. It was found through further analysis that there are significant characteristics related to specific levels of suicide risk, in which low risk tweets have a higher sentiment score; medium risk tweets prominently consist of terms that induce suicide ideation; while high risk tweets usually consist of strong and conclusive terms and have a lower sentiment score. Based on our findings, it was conclusively proven that the combination of feature extraction techniques used in our approach improves the predictive ability of the model as it helps the model understand the context of the tweet in broader perspective. Furthermore, the model takes into account the collective strengths of these features in order to synthesize and comprehend the key information conveyed by the tweet, overcoming the individual limitations of each feature extraction technique.

However, there are limitations to the performance of the model if it only considers an individual feature to determine the classification label, as the values extracted from the tweet is dependent on the choice of words used, whereby ambiguous expressions would result in misclassification. Besides that, feature extraction techniques such as PoS tagging were found to have no significant impact on the predictive ability of the model. This proves that, rather than relying solely on an individual entity, utilizing a combination of features helps the model to better understand the context of the whole tweet, which is critical towards the efficacy of the model's prediction. The collective capability of each entity forms a higher determination point that overcomes their individual limitations, resulting in better predictions. To that end, sufficient context in a tweeted post is crucial for more effective model performance, for more information is provided to the model to learn, thus yielding better prediction performance. The findings from this study contribute towards future research works towards better identification of red flags

related to suicide ideation, particularly in social media context. This creates opportunity for further extension of this study by introducing systematic intervention mechanisms in response to the tweets flagged by the model. This could be helpful in connecting the individual at risk to reliable professional help, hence leveraging suicide prevention responses with the individual at risk of suicide.

REFERENCES

[1] E. R. Kumar and N. Venkatram, "Predicting and analyzing suicidal risk behavior using rule-based approach in Twitter data," *Soft comput*, pp. 1–9, 2023.

[2] T. Zhang, K. Yang, S. Ji, and S. Ananiadou, "Emotion fusion for mental illness detection from social media: A survey," *Information Fusion*, vol. 92, pp. 231–246, 2023.

[3] R. Haque, N. Islam, M. Islam, and M. M. Ahsan, "A comparative analysis on suicidal ideation detection using NLP, machine, and deep learning," *Technologies (Basel)*, vol. 10, no. 3, p. 57, 2022.

[4] R. A. Bernert, A. M. Hilberg, R. Melia, J. P. Kim, N. H. Shah, and F. Abnousi, "Artificial intelligence and suicide prevention: a systematic review of machine learning investigations," *Int J Environ Res Public Health*, vol. 17, no. 16, p. 5929, 2020.

[5] A. Pourmand, J. Roberson, A. Caggiula, N. Monsalve, M. Rahimi, and V. Torres-Llenza, "Social media and suicide: a review of technology-based epidemiology and risk assessment," *Telemedicine and e-Health*, vol. 25, no. 10, pp. 880–888, 2019.

[6] S. E. Clark, M. C. Bledsoe, and C. J. Harrison, "The role of social media in promoting vaccine hesitancy," *Curr Opin Pediatr*, vol. 34, no. 2, pp. 156–162, 2022.

[7] J. Lee and S. Kim, "Social media advertising: The role of personal and societal norms in page like ads on Facebook," Journal of Marketing Communications, vol. 28, no. 3, pp. 329–342, Aug. 2019, doi:10.1080/13527266.2019.1658466.

[8] J. Knoll, "Advertising in social media: a review of empirical evidence," *Int J Advert*, vol. 35, no. 2, pp. 266–300, 2016.

[9] H. Ng, M. S. Jalani, T. T. V. Yap, and V. T. Goh, "Performance of Sentiment Classification on Tweets of Clothing Brands," Journal of Informatics and Web Engineering, vol. 1, no. 1, pp. 16–22, Mar. 2022, doi: 10.33093/jiwe.2022.1.1.2.

[10] A. Mbarek, S. Jamoussi, and A. Ben Hamadou, "An across online social networks profile building approach: Application to suicidal ideation detection," *Future Generation Computer Systems*, vol. 133, pp. 171–183, 2022.

[11] S. Ji, C. P. Yu, S. Fung, S. Pan, and G. Long, "Supervised learning for suicidal ideation detection in online user content," *Complexity*, vol. 2018, 2018.

[12] R. Skaik and D. Inkpen, "Suicide Ideation Estimators within Canadian Provinces using Machine Learning Tools on Social Media Text," *Journal of Advances in Information Technology Vol*, vol. 12, no. 4, 2021.

[13] F. M. Shah, F. Haque, R. U. Nur, S. Al Jahan, and Z. Mamud, "A hybridized feature extraction approach to suicidal ideation detection from social media post," in *2020 IEEE Region 10 Symposium (TENSYMP)*, 2020, pp. 985–988.

[14] S. T. Rabani, Q. R. Khan, and A. Khanday, "Detection of suicidal ideation on Twitter using machine learning & ensemble approaches," *Baghdad Science Journal*, vol. 17, no. 4, p. 1328, 2020.

[15] X. Liu et al., "Proactive Suicide Prevention Online (PSPO): Machine Identification and Crisis Management for Chinese Social Media Users With Suicidal Thoughts and Behaviors," Journal of Medical Internet Research, vol. 21, no. 5, p. e11705, May 2019, doi: 10.2196/11705.

[16] A. Mbarek, S. Jamoussi, A. Charfi, and A. Ben Hamadou, "Suicidal Profiles Detection in Twitter.," in *WEBIST*, 2019, pp. 289–296.

[17] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, "Detection of suicide ideation in social media forums using deep learning," *Algorithms*, vol. 13, no. 1, p. 7, 2019.

[18] S. Ji, S. Pan, X. Li, E. Cambria, G. Long, and Z. Huang, "Suicidal ideation detection: A review of machine learning methods and applications," *IEEE Trans Comput Soc Syst*, vol. 8, no. 1, pp. 214–226, 2020.

[19] P. Jain, K. R. Srinivas, and A. Vichare, "Depression and Suicide Analysis Using Machine Learning and NLP," in *Journal of Physics: Conference Series*, 2022, p. 12034.

[20] Y. Q. Lim, M. J. Lee, and Y. L. Loo, "Towards A Machine Learning Framework for Suicide Ideation Detection in Twitter," 2022 3rd International Conference on Artificial Intelligence and Data Sciences (AiDAS), Sep. 2022, doi: 10.1109/aidas56890.2022.9918782.

[21] B. Harmer, S. Lee, D. TvH, and A. Saadabadi, "Suicidal ideation," 2020.

[22] H.-C. Shing, S. Nair, A. Zirikly, M. Friedenberg, H. Daumé III, and P. Resnik, "Expert, crowdsourced, and machine assessment of suicide risk via online postings," in *Proceedings of the fifth workshop on computational linguistics and clinical psychology: from keyboard to clinic*, 2018, pp. 25–36.

[23] B. O'dea, S. Wan, P. J. Batterham, A. L. Calear, C. Paris, and H. Christensen, "Detecting suicidality on Twitter," *Internet Interv*, vol. 2, no. 2, pp. 183–188, 2015.

[24] R. Ahuja, A. Chug, S. Kohli, S. Gupta, and P. Ahuja, "The impact of features extraction on the sentiment analysis," *Procedia Comput Sci*, vol. 152, pp. 341–348, 2019.

[25] P. Maken, A. Gupta, and M. K. Gupta, "A study on various techniques involved in gender prediction system: a comprehensive review," *Cybernetics and Information Technologies*, vol. 19, no. 2, pp. 51–73, 2019.

[26] A. E. Aladağ, S. Muderrisoglu, N. B. Akbas, O. Zahmacioglu, and H. O. Bingol, "Detecting suicidal ideation on forums: proof-of-concept study," *J Med Internet Res*, vol. 20, no. 6, p. e9840, 2018.

[27] K. Dineva and T. Atanasova, "Systematic Look at Machine Learning Algorithms–Advantages, Disadvantages and Practical Applications," *International Multidisciplinary Scientific GeoConference: SGEM*, vol. 20, no. 2.1, pp. 317–324, 2020.

[28] A. Culotta, N. K. Ravi, and J. Cutler, "Predicting Twitter user demographics using distant supervision from website traffic data," *Journal of Artificial Intelligence Research*, vol. 55, pp. 389–408, 2016.

[29] I. I. James and V. I. Osubor, "Hostile social media harassment: A machine learning framework for filtering anti-female jokes," *Nigerian Journal of Technology*, vol. 41, no. 2, pp. 311–317, 2022.

[30] P. Kumar, P. Samanta, S. Dutta, M. Chatterjee, and D. Sarkar, "Feature based depression detection from twitter data using machine learning techniques," *Journal of Scientific Research*, vol. 66, no. 2, pp. 220–228, 2022.

[31] M. Monselise and C. C. Yang, "AI for Social Good in Healthcare: Moving Towards a Clear Framework and Evaluating Applications," in *2022 IEEE 10th International Conference on Healthcare Informatics (ICHI)*, 2022, pp. 470–471.

[32] T. Yang et al., "Fine-grained depression analysis based on Chinese micro-blog reviews," Information Processing & Management, vol. 58, no. 6, p. 102681, Nov. 2021, doi: 10.1016/j.ipm.2021.102681.

[33] S. Parrott, B. C. Britt, J. L. Hayes, and D. L. Albright, "Social media and suicide: a validation of terms to help identify suicide-related social media posts," *J Evid Based Soc Work*, vol. 17, no. 5, pp. 624–634, 2020.

[34] T. Bhardwaj, P. Gupta, A. Goyal, A. Nagpal, and V. Jha, "A Review on Suicidal Ideation Detection Based on Machine Learning and Deep Learning Techniques," in *2022 IEEE World AI IoT Congress (AIIoT)*, 2022, pp. 27–31.

[35] H. Metzler, H. Baginski, T. Niederkrotenthaler, and D. Garcia, "Detecting potentially harmful and protective suicide-related content on twitter: machine learning approach," *J Med Internet Res*, vol. 24, no. 8, p. e34705, 2022.

[36] A. L. Nobles, J. J. Glenn, K. Kowsari, B. A. Teachman, and L. E. Barnes, "Identification of imminent suicide risk among young adults using text messages," in *Proceedings of the 2018 CHI conference on human factors in computing systems*, 2018, pp. 1–11.