applied sciences

MDPI

Article

Detecting Suicidal Ideations in Online Forums with Textual and Psycholinguistic Features

Eldar Yeskuatov, Sook-Ling Chua * and Lee Kien Foo

Faculty of Computing and Informatics, Multimedia University, Persiaran Multimedia, Cyberjaya 63100, Malaysia; 1211405136@student.mmu.edu.my (E.Y.); lkfoo@mmu.edu.my (L.K.F.)
* Correspondence: slchua@mmu.edu.my; Tel.: +60-3-83125579

Abstract: Suicide is a global public health problem that takes hundreds of thousands of lives each year. The key to effective suicide prevention is early detection of suicidal ideations and timely intervention. However, several factors hinder traditional suicide risk screening methods. Primarily, the social stigma associated with suicide presents a challenge to suicidal ideation detection, as existing methods require patients to explicitly communicate their suicidal propensities. In contrast, progressively more at-risk people choose online platforms—such as Reddit—as their preferred avenues for sharing their suicidal experiences and seeking emotional support. As a result, these online platforms have become an unobtrusive source of user-generated textual data that can be used to detect suicidality with supervised machine learning and natural language processing techniques. In this paper, we proposed a suicidal ideation detection approach that combines textual and psycholinguistic features extracted from the Reddit forum. Subsequently, we selected the most informative features using the Boruta algorithm and employed four classifiers: logistic regression, naïve Bayes, support vector machines, and random forest. The naïve Bayes models trained with the combination of term frequency-inverse document frequency (TF-IDF) and National Research Council (NRC) features demonstrated the highest performance, obtaining a F1 score of 70.99%. Our experimental results illustrate that a combination of textual and psycholinguistic features yields better classification performance compared to using those features separately.

Keywords: suicidal ideation detection; machine learning; natural language processing; Reddit; psycholinguistic features

check for updates

Citation: Yeskuatov, E.; Chua, S.-L.; Foo, L.K. Detecting Suicidal Ideations in Online Forums with Textual and Psycholinguistic Features. *Appl. Sci.* **2024**, *14*, 9911. https://doi.org/10.3390/app14219911

Academic Editor: Raúl Quevedo-Blasco

Received: 9 August 2024
Revised: 17 October 2024
Accepted: 22 October 2024
Published: 29 October 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Suicide is a global phenomenon that has profound long-lasting effects on families and communities. It has been reported by the World Health Organization (WHO) that approximately 703,000 people die by suicide annually [1]. In 2019, 1 in every 100 deaths was caused by suicide. In addition to that, for each completed suicide, there are well over 20 attempts [2].

While suicide occurs among people of all ages, it is revealed to be the fourth leading cause of death in the 15–29 age bracket, preceded by road traffic accidents, tuberculosis infections, and interpersonal acts of violence [1]. Suicide is a complex issue that can be caused by a multitude of interconnected risk factors, such as mental health conditions, physical illnesses, harassment, stressful life events, childhood abuse, and a family history of suicide [3]. These risk factors are broadly categorized into three groups, such as health, environment, and personal history [4]. More recently, the social distancing measures—put in place to prevent the spread of the COVID-19 virus—further aggravated the problem [5]. This is due to the resulting isolation that has been linked to an increased risk of suicide.

Suicidal ideation refers to a wide range of thoughts and behaviors that indicate an individual's suicidal inclinations, such as extensively ruminating about taking one's own life [6]. Although they vary in their nature, intensity, and duration, suicidal ideations are

generally characterized as passive and active. Examples of passive suicidal ideations may include fleeting thoughts about committing suicide, imagining one's death, or wishing to be dead. In comparison, active suicidal ideation involves deliberately planning and preparing for an attempt, for example, by obtaining a means of suicide such as firearms or drugs. Even though passive suicidal ideations are believed to present lower risk [7], both types ought to be thoroughly assessed by clinicians, because whether active or passive, the underlying motivation is the same—terminating one's own life.

Early identification of suicidal ideations is one of the four critical strategies for effective prevention, as outlined in the WHO guidelines [8]. Other strategies also include limiting access to methods of suicide, responsible reporting of suicide cases by the media, and supporting adolescents in cultivating socio-emotional skills. By detecting suicidal ideation in its early phases, healthcare professionals can promptly intervene and provide the necessary care [9]. However, several factors hinder suicidal ideation detection, including limited access to mental health services, scarcity of trained clinicians, and fear of stigmatization [10]. The stigmatization faced by suicidal individuals prevents them from accessing social support [8] and may impede the current detection methods, such as interviews and questionnaires, as they entail explicit disclosure of suicidal propensities [11].

In recent years, progressively more at-risk people have chosen social media and online forums as their preferred avenues for sharing suicidal thoughts and seeking support, largely due to the anonymity online platforms provide [12]. As a result, there is a significant body of work that addresses suicidal ideation detection using user-generated content from online platforms. Social media sites such as Twitter, Facebook, Reddit, and Weibo have been commonly used as sources of data. Out of these, Reddit has appeared as the most notable one owing to its characteristic attributes, such as a high character limit on posts, emphasis on anonymity, and prevalence of moderated online subcommunities dedicated to various mental health topics [13].

The studies in the domain that address suicidal ideation detection with machine learning (ML) typically utilize a wide array of methods to generate features, including the bag-of-words, part-of-speech, term frequency-inverse document frequency (TF-IDF), topic modeling with latent Dirichlet allocation, sentiment analysis, linguistic inquiry and word count (LIWC), and various lexicons. The resulting features are then used as input into classification algorithms such as logistic regression (LR), support vector machines (SVM), naïve Bayes (NB), decision trees (DT), random forest (RF), and XGBoost. Most studies combine different types of features. However, it tends to lead to high dimensionality of the resulting matrix, which can be problematic. We address this concern in our paper by implementing the Boruta feature selection technique to reduce the number of features.

In addition to that, the majority of studies primarily focus on achieving higher classification performance, with few of them investigating which features indeed contribute to the effective classification of suicidal texts. This is potentially concerning, because the identification of indicative features can provide crucial insights into suicidal people's thought processes and behaviors. Although some studies perform comparative analyses of suicidal and non-suicidal texts, they are generally limited to *n*-gram analysis [14–16]. In this paper, we investigated which textual and psycholinguistic features were the most contributive to suicidal ideation detection.

Our primary aim of this paper is to present a machine learning approach for suicidal ideation detection that combines textual and psycholinguistic features extracted from Reddit data. Firstly, we extracted three types of features using TF-IDF and LIWC techniques and National Research Council (NRC) lexicons. Afterward, we applied the Boruta algorithm to select the most informative features for suicidal ideation detection. A set of four algorithms—LR, NB, SVM, and RF—were trained, and their classification performance was evaluated using 5-fold cross-validation. We also experimented with different combinations of features to determine which one resulted in the higher performance in classifying suicidal texts. The main contributions of this study are as follows:

- Identification of textual and psycholinguistic features that were the most informative for the classification of suicidal texts using the Boruta algorithm. The application of Boruta in current studies is mainly applied to either personality data or clinical assessments obtained through surveys or questionnaires [17,18].
- Experimentally determined which combination of textual and psycholinguistic features led to improved classification performance.

The article is structured as follows: Section 2 reviews related studies on suicidal ideation detection on Reddit. Section 3 describes the methodology of our proposed approach and experiment design. Section 4 presents the results. Section 5 discusses our findings, and in Section 6, the conclusions of this study are summarized.

2. Related Works

Current research approaches the task as a text classification problem by utilizing supervised ML algorithms, deep learning (DL) models, and natural language processing techniques.

For example, Aladağ et al. [19] built a ML classifier that differentiates suicidal and non-suicidal forum posts. For the dataset creation, the posts were downloaded from r/SuicideWatch, r/Depression, r/Anxiety, and r/ShowerThoughts subreddits. Then, a subset of posts was annotated under the consultation of two psychiatrists. The features extracted using TF-IDF, LIWC, and sentiment analysis were used to train four classification algorithms: ZeroR, LR, RF, and SVM. The authors of this study also applied ANOVA to reduce the number of TF-IDF features.

Bitew et al. [20] trained SVM and LR models with TF-IDF and emotion features to detect varying levels of suicide risk. The features were extracted from the UMD Reddit Suicidality Dataset created by Shing et al. [21] for the 2019 Computational Linguistics and Clinical Psychology Workshop (CLPsych 2019). The posts in this dataset are labeled as no risk, low risk, moderate risk, or severe risk.

To address the same multi-level suicide risk detection task, Rissola et al. [22] utilized TF-IDF, LIWC, Global Vectors for Word Representation (GloVe), and statistical features. The resulting features were used to train LR, SVM, and RF. This study also applied the chi-square features selection method to tackle high dimensionality.

Similarly, Shah et al. [23] focused on addressing high dimensionality in their study, exploring both filter and wrapper feature selection techniques to reduce TF-IDF, *n*-gram, and LIWC features. The researchers utilized a dataset by Ji et al. [12] for their analysis. In the study, NB, SVM, RF, and k-nearest neighbors algorithms were trained.

In addition to machine learning methods, studies often apply DL models to detect suicidal ideation. These studies frequently implement convolutional neural networks (CNNs) and long short-term memory (LSTM) techniques. However, unlike traditional machine learning, deep learning methods do not require feature engineering. Instead, they use word embeddings generated with GloVe and Word2Vec.

For instance, Ji et al. [12] addressed suicidal ideation detection with both ML and DL methods. However, in this study, both Reddit and Twitter were used as data sources. The Twitter dataset was manually annotated by researchers. The datasets were then used to train ML algorithms such as SVM, RF, Gradient Boost, and XGBoost, as well as two DL models: feed-forward neural network and LSTM. The features for the ML models were extracted with LIWC, latent Dirichlet allocation, TF-IDF, and part-of-speech techniques. The Word2Vec word embedding technique was applied to create a vector representation of textual data for training DL networks. The study demonstrated the feasibility and practicability of using ML and DL for detecting suicidal ideation and found that the performance improved when models were trained with more features.

The dataset created in [12] was also utilized by Tadesse et al. [15]. Similarly, the authors of this study compared a range of methods consisting of both ML and DL classifiers. TF-IDF, bag-of-words, and statistical characteristics were used as features when training the RF, NB, SVM, and XGBoost, while Word2Vec embeddings were used with LSTM and CNN.

The proposed ensemble comprising LSTM and CNN outperformed the baseline models. The results supported the strength of DL algorithms in suicidal ideation detection.

Similarly, the ensemble model was explored in [14]. The study used the UMD Reddit Suicidality Dataset. The authors of this study proposed a LSTM and CNN combined ensemble model to address a multiclass classification problem for detecting different levels of suicidality. The combined LSTM and CNN ensemble model enables the strengths of both methods to be leveraged. While CNN is effective at capturing local patterns and features in text, such as specific phrases indicating suicidal ideations, LSTM can capture long-term dependencies in the Reddit posts.

Gaur et al. [13] addressed the development of mapping and learning models for estimating the suicide risk severity based on forum content. The dataset consisted of r/SuicideWatch posts and was annotated by four practicing psychiatrists. The domain-specific knowledge was incorporated into the annotation process by applying the Columbia Suicide Severity Rating Scale questionnaire conventionally used by clinicians for suicide risk screening. Statistical, linguistic, and TF-IDF features were utilized to train SVM and RF. The word embeddings generated with ConceptNet were used with CNN and feed-forward neural network models. The performances of the resulting models were compared. Their results showed that CNN attained the highest performance in detecting different levels of suicidal ideations.

Yao et al. [24] aimed to detect suicidality in Reddit posts among opioid users. The data were collected from r/SuicideWatch, r/Depression, and opioid-related subreddits. As control data, the researchers used posts from subreddits not related to mental health. The dataset was annotated by crowdsourcing workers who were provided with annotation instructions. This study trained baseline ML models consisting of LR, RF, and SVM with TF-IDF features. Word embeddings from GloVe and FastText were used with CNN and recurrent neural networks. Their proposed CNN model outperformed other methods, demonstrating the viability of neural networks for the identification of suicidal ideations among opioid users.

Aldhyani et al. [25] explored XGBoost, CNN, and bidirectional LSTM models for suicidal ideation detection. The study used a publicly available unannotated dataset comprising an equal number of r/SuicideWatch posts and control posts unrelated to mental health topics. The Word2Vec embedding technique and TF-IDF were applied to generate vector representations of Reddit posts. The authors also performed a comparative analysis of suicidal and non-suicidal texts. The results of this study revealed that the CNN-LSTM combined model outperformed XGBoost at classifying suicidal texts.

A recent trend in suicidal ideation detection is the application of pre-trained transformer models, such as Bidirectional Encoder Representations from Transformers (BERTs). BERT is pre-trained on a large corpora of general text, allowing it to gain a broad understanding of natural language. These pre-trained models are then fine-tuned for detecting suicidal ideations using task-specific Reddit datasets. Studies have consistently demonstrated that transformer-based models achieve superior performance in detecting suicidal ideations.

For instance, Ambalavanan et al. [26] used the UMD Reddit Suicidality Dataset to fine-tune the BERT model for detecting varying degrees of suicide risk. Matero et al. [27] also utilized this dataset but implemented a different approach. They reduced the dimensionality of BERT embeddings using non-negative matrix factorization and combined them with lexicon-based and statistical features. Their results indicated that combining BERT with theoretical features led to improved performance in classifying texts into different suicide risk levels.

Haque et al. [28] fine-tuned multiple BERT models for the binary classification of suicidal texts and compared their performance to a LSTM baseline trained with GloVe embeddings. Their experimental results demonstrated that the BERT-based models were more effective in detecting suicidal ideation, outperforming the baseline model.

As an alternative to BERT, Nikhileswar et al. [16] employed the transformer-based Universal Sentence Encoder to generate sentence-level embeddings of Reddit posts. Their proposed model, consisting of the Universal Sentence Encoder and fully connected layers, outperformed both machine learning baselines built with SVM and XGBoost trained with TF-IDF features, as well as deep learning models comprising LSTM and CNN trained with Word2Vec.

The review of the research in this domain revealed that it is feasible to detect suicidal ideations on Reddit forums with various ML and DL techniques, as demonstrated by the aforementioned studies. However, it was also uncovered that most studies prioritize high performance, and few investigate contributive features to suicidal ideation detection. In addition to that, the studies, which compare suicidal and non-suicidal Reddit posts, are typically limited to *n*-gram analysis [14–16].

3. Materials and Methods

This study aims to develop a suicidal ideation detection model that combines textual and psycholinguistic features for detecting suicidal ideation on Reddit forums. In this section, we describe the dataset used in our study, discuss the model development process, and outline the experiment design.

3.1. Dataset

In this study, we utilized the dataset created by Aladağ et al. [19]. Each example consists of a unique identifier, post title, and body. This dataset does not include metadata such as the number of votes and comments that can be used as features. Therefore, the features were extracted from the available textual content. Any personal information such as a username was removed to protect the privacy of Reddit users. The dataset follows the binary annotation scheme and consists of 785 posts sourced from various subreddits: 175 were extracted from the r/SuicideWatch subreddit, 200 from r/Depression, 200 from r/Anxiety, and 210 from r/ShowerThoughts. Out of these extracted posts, 205 posts were labeled as suicidal and 580 posts as non-suicidal. The dataset exhibits class imbalance, which was addressed in this study through stratified sampling. This approach ensures that the training set reflects the ratio of suicidal and non-suicidal instances in the original dataset. We refrained from using undersampling, as it would result in a loss of data. Moreover, oversampling with synthetic examples was not employed to preserve the integrity of the information and the labels. Table 1 shows the distribution of the class labels in the dataset.

Table 1. Distribution of the class labels in the dataset.

Subreddit	Class		Total
	Suicidal	Non-Suicidal	
r/SuicideWatch	150	25	175
r/Depression	48	152	200
r/Anxiety	7	193	200
r/ShowerThoughts	0	210	210
Total	205	580	

This dataset was chosen since it was analyzed and annotated by experts in the mental health domain. Specifically, the annotation was carried out by two psychiatrists: The expert annotations ensure that every post labeled as suicidal genuinely reflects suicidal language. This, in turn, supports the clinical validity of the resulting models and strengthens the potential findings about the traits and behaviors of individuals with suicidal ideations.

r/SuicideWatch is a forum where users share their suicidal thoughts and experiences and seek support. The researchers frequently use this subreddit to collect suicidal texts [9]. It is often regarded as a reliable and genuine source of suicide-related posts, since r/SuicideWatch is managed by moderators who remove any unrelated posts or any

content violating community guidelines, e.g., spam, abuse, or criticism. R/Depression and r/Anxiety are subcommunities of people discussing depression and anxiety, respectively, while the r/ShowerThoughts subforum is not related to mental health. On r/ShowerThoughts, the users share miscellaneous thoughts and insights. Table 2 exemplifies the snippets of posts from both classes.

Table 2. Snippets of posts in suicidal and non-suicidal classes.

Suicidal	Non-Suicidal
I want to just not wake up most nights. If I had the chance, I'd indefinitely kill myself.	I know I [expletive] up, I can't stop beating myself up. I'm not suicidal but I don't like the thoughts going through my mind right now. I just need someone to talk to.
I planned to kill myself yesterday. I don't want to leave my son without a father, but it hurts too much to keep going.	My closest friend has been saying she's going to commit suicide for several weeks now. I need some way to help her through this, I don't know what to do.
I've lived with depression and suicidal thoughts since I was fourteen. I'm now in my late thirties. I'm still suicidal and still suffer from depression.	I just now started watching The Office and was shocked at how cruel some of the pranks are at Dwight's expense. Jim is the kind of bully who gets away with it because he's cute and it's [expletive].

3.2. Data Preprocessing

The data extracted from online platforms such as Reddit tend to contain noise and artefacts that can disrupt machine learning algorithms. The noise can be in the form of HTML tags, punctuations, special characters, URLs, and emails. Also, Reddit data comprise raw unstructured text that needs to be standardized. Therefore, data preprocessing is an essential step that prepares data for feature extraction. In our work, we applied the following data preprocessing steps.

3.2.1. Concatenating Post Title and Body

A Reddit post consists of a title and a body. During this step, we concatenated the two because the title lengths vary in the dataset. It can range from one word to several sentences. In a few instances, the titles were non-descriptive or did not contain any words. Therefore, the post title on its own is not reliable for extracting features and should be added to the body.

3.2.2. Expanding Contractions

During this step, the contractions were expanded. For example, phrases such as "won't" and "didn't" were converted to "will not" and "did not". The reason we included this step was to facilitate the better detection of negation and verb tenses. We utilized the Contractions (version 0.1.73) Python library to perform this step [29].

3.2.3. Data Cleaning

During this step, we removed numbers and any elements that did not carry semantic meaning, including URLs, HTML tags, punctuations, and line break character sequences. We also removed any emojis present in the data instead of converting them into corresponding textual descriptions. According to current studies on the use of emojis, people tend to use polite and smiling emojis in a negative context [30]. This kind of discordance between emotions experienced and expressed with emojis could potentially affect the classification model. To carry out data cleaning, we used Python's built-in regular expression (version 3.10.10) library.

3.2.4. Lemmatization

We conducted lemmatization to convert different inflections of words into a common root, lemma. Lemmatization allows semantically similar words to be grouped. For example, the words "drives", "drove", and "driven" are converted to "drive". This step helps reduce the number of textual features. We used the SpaCy (version 3.5.1) open-source Python library for this task [31]. It is worth noting that lemmatization was performed as a prerequisite step for generating TF-IDF and NRC features only but not for LIWC. LIWC contains features pertinent to past, present, and future focuses. If we were to convert all forms of a verb into the present tense, the LIWC analysis tool would not be able to capture these features.

3.2.5. Tokenization

We applied tokenization to break down the long passages of text into tokens. In our work, tokenization was done on a word level. Essentially, one word represents a single token. We chose word-level tokenization, because NRC feature values are calculated based on the matches of individual words from a post with the lexicons representing different emotions. The same is true for LIWC features. However, the LIWC tool performs tokenization behind the scenes and thus does not require the input data to be tokenized. We used Python's built-in string (version 3.10.10) library for splitting text into individual words.

3.3. Feature Extraction

After the data preprocessing stage, we proceeded to extract features that served as input to a classifier algorithm. We primarily focused on extracting features from the textual content of posts, as the datasets do not contain any statistical metadata, such as timestamps or the number of comments and votes. The following subsection details the feature extraction stage.

3.3.1. Term Frequency-Inverse Document Frequency

We applied TF-IDF to create a vector representation of preprocessed textual Reddit data. In the large body of text, the terms that appear very frequently (e.g., "is", "are", "the", and "a") have low semantic importance to the content of the document. TF-IDF considers how frequently the term appears in the corpus and penalizes the most frequent terms, giving rarer terms higher importance. Concerning the Reddit dataset, a document refers to an individual post. The TF-IDF value of a term t in a given document d is the product of term frequency (tf) and inverse document frequency (idf) and is calculated as

$$tf - idf(t,d) = tf(t,d) \times idf(t) \tag{1}$$

where $tf(t,d)$ is the number of times a term t appears in a given document d. The $idf(t)$ component is calculated as

$$idf(t) = \log_e \frac{1+n}{1+df(t)} + 1 \tag{2}$$

where n is the total number of documents in the corpus (i.e., posts in the dataset), and $df(t)$ is the number of documents d that contains the specific term t. The value is calculated for all terms in the corpus. The output of this process is a sparse matrix, where each row is a vector of floating point values that numerically represents a single post. Each resulting vector is then normalized by applying the Euclidian norm. Each column of the matrix represents a single term that serves as a feature. Figure 1 shows an example of the resulting matrix. The values in this example are arbitrary. To generate TF-IDF features, we utilized the TfidfVectorizer package from the scikit-learn (1.5.1) library [32]. The resulting matrix contained, on average, 5902 features.

	$term_1$	$term_2$	$term_3$	...	$term_{m-1}$	$term_m$
$post_1$	0.1887	0.0000	0.1033	...	0.0000	0.0674
$post_2$	0.2026	0.0000	0.1110	...	0.0375	0.0495
$post_3$	0.0164	0.0000	0.0000	...	0.0809	0.0000
...	...	...	...	...	...	...
$post_{n-1}$	0.0000	0.1206	0.0000	...	0.6142	0.0311
$post_n$	0.1790	0.0000	0.4166	...	0.4166	0.0000

Figure 1. Example of the TF-IDF matrix.

3.3.2. Linguistic Inquiry and Word Count

We applied LIWC to extract psycholinguistic features from Reddit posts. LIWC-v.22 is a software tool that allows us to gain insight into a person's psychological state based on language use [33]. At the core of this tool is a dictionary of over 12,000 word stems, words, phrases, and select emoticons. The dictionary is divided into 118 categories, each intended to capture various psychosocial constructs. LIWC was chosen to generate psycholinguistic features due to its ability to capture the psychological properties of language (e.g., negative affect), making it particularly suitable for suicidal ideation detection. The effectiveness of this feature extraction method has been validated in prior research [12,19], further supporting its use in this study.

LIWC compares each term in the post against the list of dictionary words and calculates the percentage of total terms that match the specific category. For example, if LIWC analyzes a single Reddit post containing 100 words and finds that 5 of those words are related to sadness, the post scores 5% in the *emo_sad* category. Most LIWC features are calculated as percentages, except for four summary features, including analytical thinking, clout, authenticity, and emotional tone. They are calculated using proprietary algorithms not disclosed in LIWC documentation. In our work, we used the LIWC-22 (Pennebaker Conglomerates Inc., Austin, TX, USA) version of the tool [34]. We extracted 108 features, as we excluded features related to punctuation and emoticons.

3.3.3. National Research Council Lexicons

NRC lexicons are a list of word association lexicons that are primarily used to analyze sentiments and emotions in texts. We utilized the NRC lexicons to incorporate emotion-specific features into our analysis. By employing features that convey the intensity of specific emotions expressed in text, we sought to enhance the differentiation between suicidal and non-suicidal texts. In this paper, two NRC lexicons were utilized: the Emotion Intensity Lexicon (EIL) [35] and the Valence, Arousal, and Dominance lexicon (VAD) [36].

The EIL consists of 9829 words with associated intensity scores for eight fundamental emotions, including anger, anticipation, disgust, fear, joy, sadness, surprise, and trust. These emotion categories are based on the Wheel of Emotions psychological theory [37]. For a given term and emotion, the score can range from 0 to 1. If the score is 1 or very close to 1, then the term conveys the highest expression of the said emotion. Conversely, if the score is 0 or close to it, then the term conveys the lowest expression.

The VAD lexicon consists of 23,342 words and their associated scores related to the valence, arousal, and dominance dimensions that describe emotions. Similar to EIL, for a given term and a dimension, the score can range from 0 to 1, with 0 representing the lowest and 1 representing the highest expression of a particular dimension. Valence refers to a degree to which the emotion is negative or positive, e.g., a high valence indicates positive emotions. Arousal refers to the intensity of the emotion. Dominance represents the level of control the person feels. It ranges from being submissive to dominant. For example, fear and anger are both negative, but anger is a more dominant emotion compared to fear.

Two features are generated for each psycholinguistic category from both lexicons: average lexicon value *avgLexVal* and lexicon ratio *lexRatio*. For example, for a fear category

from EIL, there are two features called *fear_inten_avgLexVal* and *fear_inten_lexRation*. Every term from a post is compared to lexicon words from a particular emotion category. *AvgLexVal* is calculated as the average score of all matching words. For example, if a post contains the words "torture", "kill", and "suicide", their associated scores from the fear category ($score_{torture}$ = 0.984, $score_{kill}$ = 0.962, and $score_{suicide}$ = 0.879) will be averaged, and the resulting *fear_inten_avgLexVal* will be 0.942. *LexRatio* is calculated as the number of matching words divided by the total number of words in the post. For example, the *valence_vad_lexRation* score of 0.15 signifies that 15% of the words in the post matched with words from the valence category. To generate NRC features, we used the Emotion Dynamics (version 1) Python package [38]. As the outcome of this step, we extracted 22 NRC features.

3.4. Feature Selection

The resulting matrices of features consisted of a large number of dimensions. High dimensionality can pose difficulties, as the processing of such large matrices is more computationally costly and prone to result in longer training times. In addition to that, it can lead to suboptimal classification performance due to the presence of numerous irrelevant features in the dataset. To address this issue, we applied the Boruta algorithm for feature selection.

Boruta is a wrapper feature selection method built on a random forest classifier [39]. First, the algorithm creates shadow features, which are randomly permutated copies of the original features. The shadow features are then concatenated to the original matrix. The final matrix has twice the number of features. Then, the random forest classifier is trained with the extended dataset, and the Gini importance—feature importance measure based on Gini impurity—is computed for all the features. The original feature's importance is compared to the highest importance calculated for the shadow features. If the original feature is less important than the most important randomized feature, it is removed. This process is iterated for a specified number of times or until all features are confirmed. The default number of iterations is 20. For our study, we implemented the Boruta algorithm using the BorutaPy (version 0.4.3) Python package [40]. The number of iterations was set to 100. We picked a higher number of iterations because the TF-IDF matrix consists of 5902 features, and it requires more iterations to select the informative features.

3.5. Model Training

Suicidal ideation detection is a classification problem that we approached with supervised machine learning algorithms. During this step, the generated features were used to train the classification. Although the underlying operation of algorithms varies, the goal is the same: to find the function F that maps the relationship between input features $\underset{x}{\rightarrow}$ and a target class y while minimizing the prediction error. We trained four classifiers: logistic regression, naïve Bayes, support vector machines, and random forest.

3.6. Model Evaluation

To evaluate the classification performance of the selected algorithms, we applied stratified 5-fold cross-validation. The cross-validation procedure provided five different splits of training and testing sets. In each split, 80% of the data (628 posts) was used for training, and 20% (157 posts) was used for testing. The stratified sampling ensures that training and testing sets have an equal ratio of suicidal posts. The performance was measured with accuracy, precision, recall, and F1 score. These metrics were computed from the confusion matrix.

Accuracy is defined as the fraction of correctly classified examples out of all classifications. Precision represents the fraction of correctly classified positive examples out of all examples classified as suicidal. Recall is the ratio of correctly classified positive examples to the total number of positive examples in the dataset. F1 score is the harmonic mean of precision and recall.

3.7. Experiment Design

In this study, we carried out three experiments. All experiments were conducted with different compositions of the input features.

Experiment 1 evaluated the performance of models trained with different types of features. In this experiment, separate sets of four algorithms were trained with TF-IDF, LIWC, and NRC features, producing 12 models in total. The TF-IDF matrix represents the textual features, whereas the LIWC and NRC matrices represent the psycholinguistic features. The goal of this experiment was to determine which individual type of features had the best predictive capability.

Experiment 2 evaluated the performance of the classifiers when feature selection was applied. TF-IDF, LIWC, and NRC features selected with Boruta were used to train separate sets of classifiers, resulting in 12 models. The purpose of this experiment was to ascertain whether models trained with reduced features had improved classification performance.

Experiment 3 evaluated the performance of models trained with different combinations of textual and psycholinguistic features. Given three types of features, we suggest four combinations: TF-IDF+LIWC, TF-IDF+NRC, LIWC+NRC, and TF-IDF+LIWC+NRC. Training four algorithms with four possible combinations of features brings about 16 classifier models. The goal of this experiment was to determine whether combining textual and psycholinguistic features resulted in better performance.

4. Results

4.1. Classification Performance

4.1.1. Experiment 1 (Individual Features)

Figure 2 shows the classification performance of models trained with individual features. The values reported here are the average of the five-fold cross-validation. Each bar chart within the figure illustrates the accuracy, precision, recall, and F1 score metrics achieved by specific algorithms, with feature types denoted using different colors.

Closer inspection of the accuracy scores reveals that algorithms trained with TF-IDF, LIWC, and NRC features demonstrate comparable accuracy. Particularly, LR, RF, and SVM achieved accuracy scores between 75% and 80%, and NB achieved approximately 60% accuracy across all feature types. Notably, in the LR and RF models, training with LIWC features resulted in higher accuracy, with RF obtaining 82.80%.

The distinctions in performance among the TF-IDF-, LIWC-, and NRC-trained models became more evident upon further analysis of the precision and recall metrics. LR, RF, and SVM trained with TF-IDF exhibited higher precision compared to the models utilizing other features. In contrast, TF-IDF-trained NB resulted in the lowest precision at 33.04%.

Training with LIWC features yielded higher recall scores across all algorithms compared to other feature types. However, TF-IDF and NRC showed mixed results, with TF-IDF-trained RF recording the lowest recall of 17.07%, and SVM trained on NRC obtaining 21.95%. RF trained with TF-IDF achieved the highest precision (86.99%) while recording the lowest recall score. Conversely, NB trained with LIWC achieved the highest recall (83.90%) but with lower precision at 38.61%. These two cases demonstrate the precision–recall trade-off.

This trade-off is further highlighted when analyzing the overall trends. Algorithms with higher precision (LR, RF, and SVM) tend to exhibit lower recall, suggesting a higher number of false-negative predictions. In contrast, the NB algorithm, which scored the lowest precision with all types of features, had the highest recall.

Considering this trade-off, relying solely on precision or recall to determine the best-performing model might be challenging. Thus, assessing the F1 score, a harmonic mean of precision and recall, provides a more comprehensive overview. Comparing F1 scores across models trained on various features, it becomes more apparent that models using LIWC features generally outperform those using TF-IDF or NRC. Specifically, RF trained with TF-IDF and SVM trained with NRC demonstrated the lowest F1 scores (28.12% and 31.17%, respectively), whereas LR trained with LIWC achieved the highest F1 score at 59.17%.

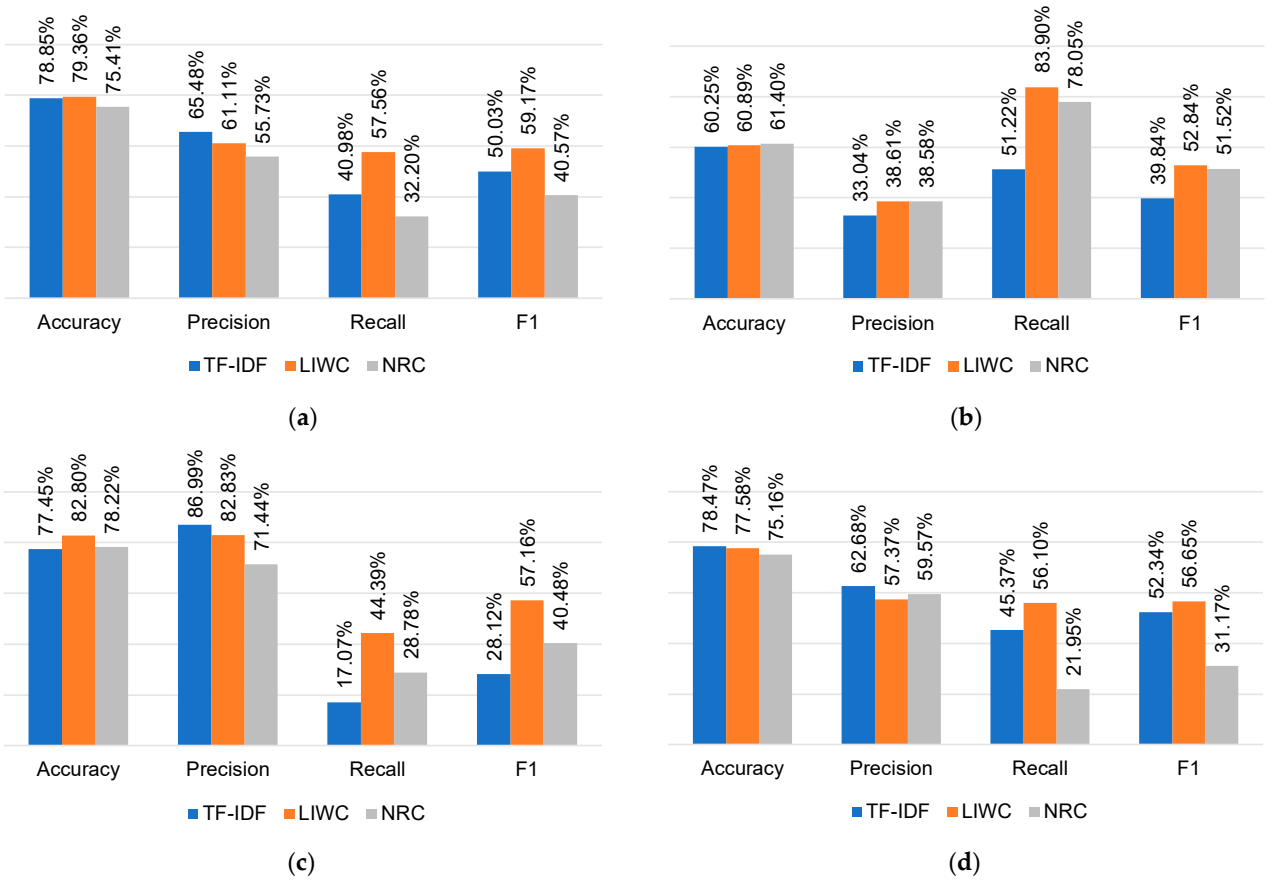


Figure 2. Performance scores of models trained with different types of features: (**a**) logistic regression; (**b**) naïve Bayes; (**c**) random forest; (**d**) support vector machines.

4.1.2. Experiment 2 (Feature Selection)

In Experiment 2, we investigated the impact of Boruta feature selection on the classification performance. The aim was to ascertain whether models trained with selected features achieved higher performance compared to models trained with all features. Figure 3 illustrates a comparison between the mean performance metrics of the models trained using all features and those trained with Boruta-selected features. Each subplot within the figure represents a distinct feature type.

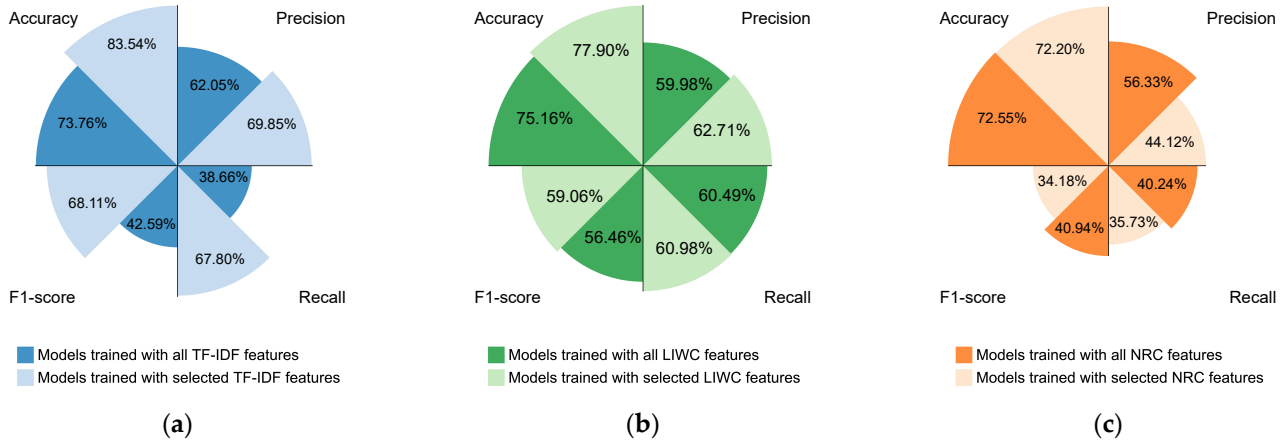


Figure 3. Average performance scores of the models trained with different types of features: (**a**) TF-IDF; (**b**) LIWC; (**c**) NRC.

As shown in Figure 3a, when feature selection was applied to TF-IDF features, it resulted in an increased performance across all metrics. Specifically, the average accuracy increased from 73.76% to 83.54% and precision from 62.05% to 69.85%. In particular, the highest increments were observed in recall and F1 scores. Recall increased by 29.14%, from 38.66% to 67.80%, while F1 score grew by 25.52% from 42.59% to 68.11%. Out of all TF-IDF-trained models, RF experienced the largest F1 score gain of 37.68% from 28.12% to 65.80%. Boruta reduced the TF-IDF features from 5902 to 32, accounting for about 0.54% of the original features. Moreover, feature selection led to a 14-fold reduction in the TF-IDF-trained models' average training time.

Similarly, when feature selection is applied to LIWC features, improvements in accuracy, precision, recall, and F1 score were observed, although to a lesser extent, as shown in Figure 3b. The average accuracy increased from 75.16% to 77.90%, precision from 59.98% to 62.71%, recall from 60.49% to 60.98%, and F1 score from 56.46% to 59.06%. Among the LIWC-trained models, NB showed the highest F1 score increase of 5.99%. The number of LIWC features was reduced from 108 to 22 (approximately 20% of the original features), while the LIWC-trained models' training time decreased by 2.43-fold.

In contrast, applying feature selection to NRC features led to a decrease in all performance metrics, as illustrated in Figure 3c. There was a slight decline in accuracy by 0.35%. However, a more prominent reduction was observed in precision, decreasing by 12.21% from 56.33% to 44.12%. Recall and F1 score dropped by 4.51% and 6.76%, respectively. Boruta reduced the NRC features by half, from 22 to 11. The average training time decreased by 20% from 53.65 ms to 43.02 ms. In summary, Boruta feature selection improved the classification performance of models trained with TF-IDF and LIWC features but resulted in a decreased performance among NRC-trained models.

4.1.3. Experiment 3 (Combined Features)

The aim of this experiment was to examine whether combining textual and psycholinguistic features results in improved performance. Various types of Boruta-selected features were combined in this experiment. Figure 4 shows the performance scores of models trained using different feature combinations. Each group bar plot in the figure represents the performance of a specific algorithm, with feature combinations denoted using different colors.

Upon closer examination of the accuracy metrics, it was observed that all the algorithms attained their lowest accuracy when trained with a combination comprising solely psycholinguistic LIWC and NRC features. Notably, NB trained with LIWC+NRC recorded the lowest accuracy at 68.41%.

Conversely, the models that utilized some combination of textual and psycholinguistic features achieved the highest accuracy. For instance, RF achieved the highest accuracy of 85.35% when trained with TF-IDF+LIWC and TF-IDF+LIWC+NRC combinations. Additionally, NB and SVM recorded the highest accuracy of 81.78% and 84.46%, respectively, when the TF-IDF+NRC feature set was used for training.

An analysis of the precision metrics revealed a similar trend. Models trained with LIWC+NRC feature combinations tended to exhibit lower precision scores. Specifically, NB achieved 44.77% precision when the LIWC+NRC combination was used as the input features. However, combining textual and psycholinguistic features led to higher precision. Particularly, NB, RF, and SVM attained the highest precision scores of 60.92%, 78.12%, and 71.84%, respectively, with TF-IDF+NRC, while LR achieved 73.98% with TF-IDF+LIWC.

In terms of recall, better performance was achieved with a combination of textual and psycholinguistic features. For example, LR and SVM achieved the highest recall (66.83% and 67.32%, respectively) when the TF-IDF+NRC feature combination was used for training. In addition, RF attained its highest recall of 64.39% with TF-IDF+LIWC, and NB recorded an 89.27% recall when all three types of features were combined.

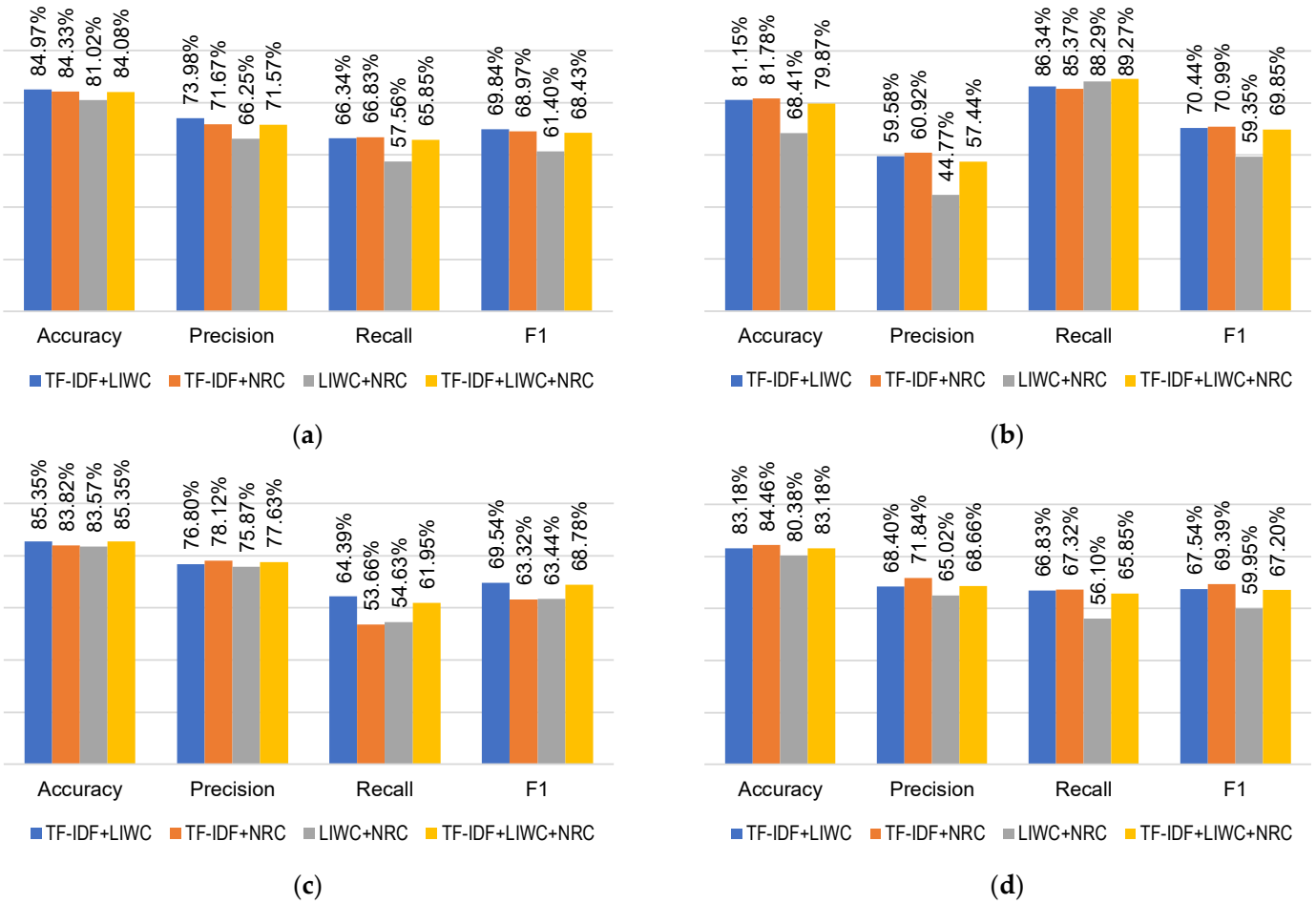


Figure 4. Performance scores of models trained with different combinations of selected features: (**a**) logistic regression; (**b**) naïve Bayes; (**c**) random forest; (**d**) support vector machines.

In contrast, combining LIWC and NRC resulted in the lowest recall for LR and SVM (57.56% and 56.10%, respectively). It is noteworthy that NB tends to exhibit higher recall and lower precision than the other algorithms.

Further inspection of F1 scores also indicates better performance when the feature set includes both textual and psycholinguistic features. Specifically, LR and RF achieved the highest F1 scores of 69.84% and 69.54%, respectively, when combining TF-IDF and LIWC features. NB and SVM performed best when combining TF-IDF and NRC, achieving F1 scores of 70.99% and 69.39%, respectively. On the contrary, when NB was trained on LIWC+NRC features (both psycholinguistic), it recorded the lowest F1 score of 59.35%.

From Figure 5, it becomes evident that combining textual and psycholinguistic features tends to result in improved classification performance compared to using these features separately. For instance, RF recorded a 65.80% F1 score with only TF-IDF features and 61.76% with only LIWC features. However, combining these features led to a higher F1 score of 69.54% for RF.

These results highlight the significance of feature combination, especially when utilizing both psycholinguistic and textual features, as it leads to improved classification performance.

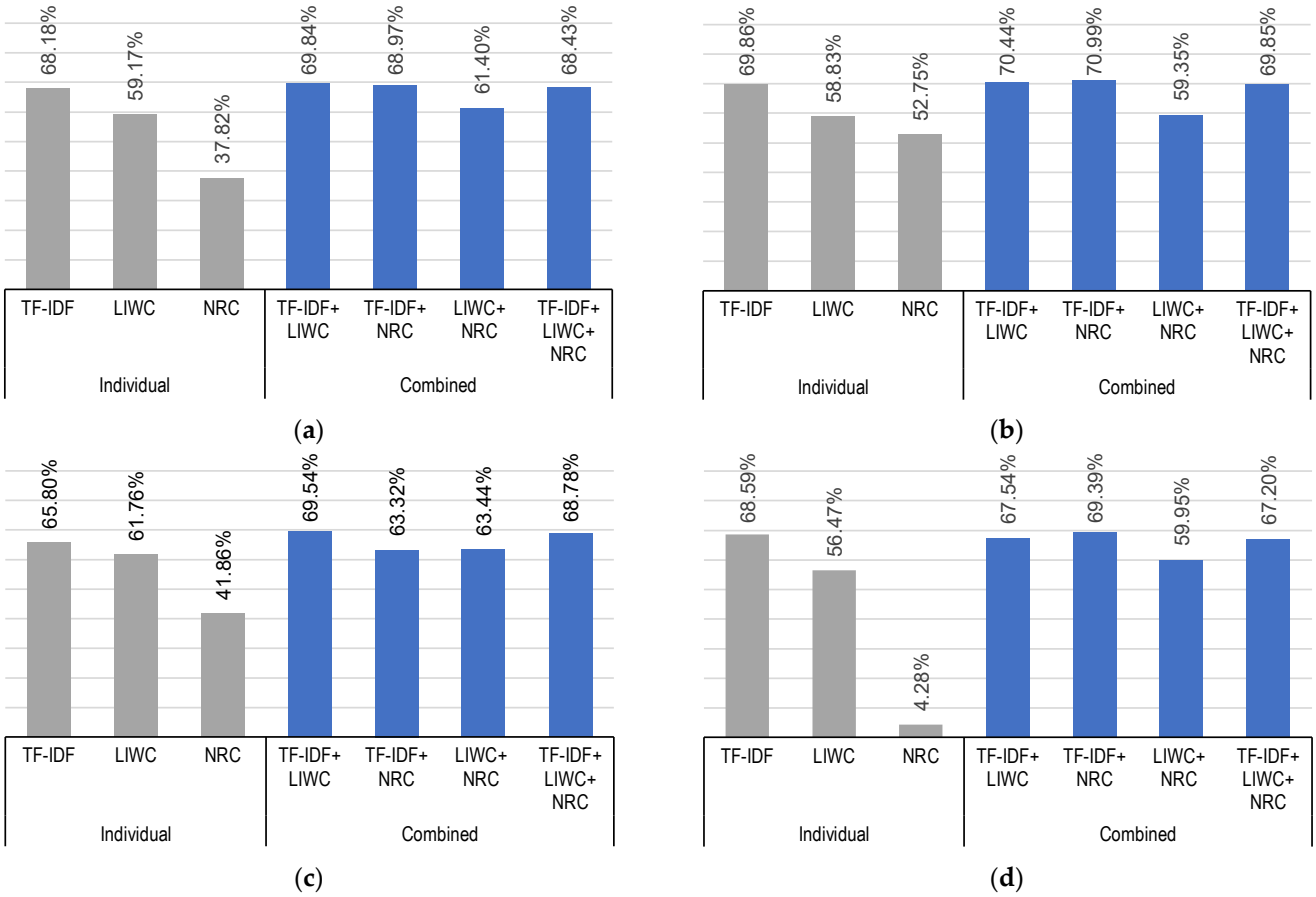


Figure 5. F1 scores of models trained with different combinations of selected features: (**a**) logistic regression; (**b**) naïve Bayes; (**c**) random forest; (**d**) support vector machines.

4.2. Selected Set of Informative Features

The features selected with the Boruta algorithm can provide insight into the way suicidal individuals communicate. As a result of the feature selection process, we extracted the list of features that have the most contribution to suicidal ideation detection. The feature selection was performed on each five-fold cross-validation split's training set. This step yielded five different sets of selected features. We identified the features that were selected in all splits by finding the intersection of five sets of extracted features. As a result of this step, we identified the 24 most informative TF-IDF features, as well as the 14 most informative LIWC features. We did not perform this process with NRC, since feature selection decreased the classification performance.

From the informative TF-IDF features, we observed that suicidal posts scored higher in features pertaining to personal pronouns such as *myself* and *my*. Suicidal texts also contain more death-related words represented by TF-IDF features *kill*, *die*, *suicide*, and *suicidal*, as well as words indicating intention and action under features *want*, *plan*, *do*, and *end*.

A closer inspection of LIWC features and their average scores revealed that suicidal texts scored higher in the first-person singular pronouns category, which also reflects the trend discovered with TF-IDF features. Other noteworthy LIWC categories where suicidal texts scored higher include *want*, *death*, *friend*, *conflict*, *allnone*, and *authenticity*. However, they had lower scores under the *clout* category. The interpretation of these results is discussed further in Section 5.

5. Discussion

5.1. Classification

5.1.1. Experiment 1 (Individual Features)

From the first experiment, we observed that models trained with TF-IDF and NRC features did not perform well, since their average F1 scores were lower than 50%. The suboptimal performance of TF-IDF-trained models could potentially be attributed to the phenomenon described as the curse of dimensionality. A high number of TF-IDF features makes the data sparse and subsequently impedes the model's ability to generalize. In the training dataset consisting of 628 examples, there are, on average, 5902 TF-IDF features. A possible explanation for the low F1 score among models trained with NRC features is the fact that NRC features correspond to only 11 psychological constructs. This is likely not sufficient for the model to effectively capture the differences between classes. Another type of psycholinguistic feature we tested was LIWC. The use of LIWC features generally resulted in better classification performance arguably, because it utilizes a more extensive set of dictionaries that correspond to a wider range of psychological processes and linguistic dimensions. This finding supports the effectiveness of LIWC at capturing underlying linguistic signals related to distress, emotional traits, and cognitive processes that indicate suicidal ideations. Consequently, it highlights the importance of incorporating psychologically grounded tools into the development of suicidal ideation detection models.

5.1.2. Experiment 2 (Feature Selection)

In the second experiment, we addressed the issue of high dimensionality. Having applied Boruta feature selection, we observed the highest improvement in classification among the models trained with TF-IDF features. This finding further supports our suggestion that the subpar performance of TF-IDF-trained models in the first experiment might have resulted from high dimensionality. The feature selection produced promising results, as we were able to obtain an approximately 25% increase in F1 score while retaining only a fraction of the original TF-IDF features. A similar trend can be observed among the models trained with LIWC features but to a lesser degree. However, the feature selection led to a declined performance of the models trained with NRC features. This outcome was not unexpected, as the number of original NRC dimensions was not large. These findings suggest that Boruta feature selection improves the classification performance when the original training set comprises a large number of dimensions. The significant improvements in the TF-IDF-trained models' performance highlight the importance of dimensionality reduction in suicidal ideation detection.

5.1.3. Experiment 3 (Combined Features)

The results of the third experiment show that combining textual features with either type of psycholinguistic feature resulted in a higher F1 score compared to using those features separately. The NB models achieved the highest classification performance in terms of F1 score and recall with this feature combination. This finding suggests that models utilizing both psycholinguistic and textual features potentially detect suicidal ideations among online forum users better than models that solely rely on textual or psycholinguistic features.

5.2. Informative Features

During the feature selection process, we were able to identify the features that had the biggest contribution to suicidal ideation detection. It is worth noting that most LIWC categories are organized hierarchically. LIWC's *linguistic* category and its subcategories *i* and *negate* are among the most informative features.

In particular, suicidal texts, on average, contain 4.24% more first-person singular pronouns (LIWC feature *i*) compared to non-suicidal texts. Examples of such pronouns include "I", "me", "myself", and "my". This trend is also reflected in TF-IDF features, where features *myself* and *my* have higher average values in suicidal texts. This finding

aligns with the existing literature [12]. Arguably, individuals with suicidal tendencies use more first-person singular pronouns as they express their psychological distress, as well as in the context of inflicting harm on themselves or ending their lives.

The suicidal texts also scored higher in the LIWC categories related to negations (e.g., "no", "not", "never", and "nothing), as well as all-or-none thinking that includes the words "all", "never", and "always". Regarding TF-IDF, this tendency can be observed in features *anymore*, *not*, and *no*. This is consistent with the conclusions of prior research [41]. The dominance of these features in suicidal texts potentially suggests that suicidal individuals exhibit all-or-nothing thinking when they communicate on suicide support forums. All-or-nothing thinking (also known as polarized thinking) is a type of exaggerated negative thought patterns called cognitive distortions [42].

Another notable pair of LIWC features that are prevalent in suicidal texts are the categories *friend* and *conflict* that pertain to social behavior. The *friend* category includes terms such as "boyfriend" and "girlfriend", while *conflict* includes "fight" and "attack". The similar pattern can be seen in TF-IDF features *friend* and *family*. The prevalence of these features possibly indicates that suicidal people are discussing relationship issues on Reddit suicide forums. This is quite plausible, as relationship conflict is known to be one of the key suicide risk factors [4].

As expected, suicidal texts score higher in the *want* and *death* LIWC categories and similar TF-IDF features such as *kill*, *end*, *life*, *die*, *want*, and *suicide*. This combination of features appears more frequently, likely because at-risk users express their desire to end their life.

Interestingly, the non-suicidal posts contain more words related to anxiety. We believe this is because about 33.27% of the non-suicidal class consists of posts from the r/Anxiety subreddit. It is also worth noting that the average LIWC *clout* score was 16.83% lower in suicidal texts. This score describes the social status, confidence, and sense of power that the person demonstrates in their writing [34]. This suggests that suicidal individuals exhibit more powerlessness compared to their non-suicidal counterparts when communicating online.

Robustness of Informative Features

To evaluate the robustness of our approach, we conducted additional testing on a separate, much larger dataset using the informative psycholinguistic and textual features identified in this study. The new dataset, consisting of 187,943 Reddit posts, initially included 300 TF-IDF and 108 LIWC features [19]. We applied 24 TF-IDF and 14 LIWC features identified in Experiment 2 to this larger dataset and compared the results. The same set of four algorithms was trained: logistic regression, naïve Bayes, random forest, and support vector machine.

Our findings show that the performance of the models trained with our selected features was comparable to that of models using the original feature set. Specifically, the models trained with the original features achieved an average F1 score of 97.08% compared to 95.29% for the models trained with our selected features. This represents a 1.79% decrease in F1 scores. A similar decrease of 1.84% was observed in the average accuracy scores: models trained with all features achieved 96.95%, while those trained with the selected features reached 95.11%.

However, while the models trained with selected features achieved slightly lower performance, they utilized only 9.31% of the original features: 38 Boruta-selected features compared to 408 original features. This reduction in features resulted in an approximately 18-fold decrease in the average training time. In particular, the models using our informative features completed training in 2.22 s, whereas the models trained with the original features took an average of 39.29 s to train. The most significant reduction in training time—from 128 s to 0.52 s—was observed in the SVM model.

These results support the robustness and efficiency of our approach. They demonstrate that the features we identified in this study are not only effective on expert-labeled datasets but also generalize well to a much larger Reddit dataset.

5.3. Limitations

The major limitation of this study stems from data scarcity. To strengthen the clinical validity of the models' predictions and their interpretations, as well as the obtained insights into suicidal users' communication idiosyncrasies, we chose to use the dataset that was manually annotated by experts in the mental health domain. In particular, utilizing a dataset by [19] where each example was analyzed and labeled by two practicing psychiatrists offers an assurance that every post from the suicidal class indeed represents suicidal language. However, this laborious process of manual annotation often means that the resulting datasets are fairly small. As a result, the limited number of training examples restricted the range of methods that could be explored in this study for suicidal ideation detection. Notably, the current psychiatrist-annotated dataset (785 examples) was insufficient for training deep learning methods such as recurrent and convolutional neural networks or more recent transformer-based models, which typically require tens of thousands of examples to learn from data effectively.

The alternative is to use large datasets created through web scraping on various Reddit forums. In this case, the annotations would be based on the community affiliation. Specifically, all posts from suicide-related forums are assumed to represent suicidal language, whereas the posts representing general language would be sourced from other forums unrelated to suicide or mental health. This approach to data collection can facilitate the use of more complex deep learning methods. However, since the labels are not verified by experts, it potentially limits the interpretability of the results, as well as the clinical validity of the predictions.

5.4. Practical Implications and Ethical Considerations

Detecting suicidal ideation on social media forums offers practical benefits. Automated detection tools could potentially identify at-risk individuals who might not otherwise seek help due to the fear of social stigma associated with mental health issues. Consequently, these online detection tools could enable more timely interventions. These systems could alleviate the burden on healthcare providers by flagging concerning posts for further review by a mental health professional, allowing for a more efficient allocation of resources. In addition to that, automated suicidal ideation detection tools could address the issue of fragmented care for existing patients, which entails long gaps between assessments. These tools could assist mental health professionals in monitoring patients between clinical interactions by analyzing patients' online activity and alerting if their mental health deteriorates.

However, this approach also raises ethical concerns. First, the protection of privacy during model development is crucial. It is important to anonymize the data used for training suicidal ideation detection models to ensure no identifying information is included. Second, after deployment, users should be informed that their online data may be analyzed, ensuring transparency and fair practices. It is also essential to provide users with an option to opt out in case they do not want their activity monitored. There is also a risk of false positives when the user is incorrectly identified as experiencing suicidal ideations. This could lead to unnecessary interventions or stigmatization. To mitigate this, it is important to avoid overreliance on automated systems, and their outputs should always be verified by qualified professionals.

In summary, the implementation of such online detection systems could add value, but the individual privacy, consent, and intervention guidelines should be carefully considered to ensure ethical use and prevent potential misuse.

5.5. Future Directions

A further study could be focused on engineering suicide-specific features by incorporating suicide lexicons. Such features could enable ML models to better differentiate suicidal posts from posts related to other mental disorders. This, in turn, could help address the issue of overlapping edge cases.

Another direction the study could take is to apply transfer learning. Transfer learning allows to leverage the power of large pre-trained language models by fine-tuning them with task-specific training datasets. These fine-tuned models could potentially learn more complex patterns in the Reddit data, consequently improving the classification performance. However, in order to effectively fine-tune complex pre-trained models, a larger dataset would be required.

6. Conclusions

In this paper, we presented a suicidal ideation detection model that combines textual and psycholinguistic features. We extracted TF-IDF, LIWC, and NRC features and applied the Boruta algorithm to find the subset of the most informative features for suicidal ideation detection. We experimentally illustrated that the combination of textual and psycholinguistic features yields better classification performance compared to using those features separately. NB models trained with TF-IDF+NRC and TF-IDF+LIWC feature combinations achieved the highest F1 scores of 70.99% and 70.44%, respectively. Finally, we analyzed the most informative features and discussed study limitations and future directions. The findings of this study contribute new insight and understanding to the science of detecting suicidal ideations on online forums.

Author Contributions: Conceptualization, S.-L.C. and L.K.F.; methodology, E.Y. and S.-L.C.; validation, E.Y., S.-L.C. and L.K.F.; formal analysis, E.Y., S.-L.C. and L.K.F.; data curation, E.Y.; writing—original draft preparation, E.Y.; writing—review and editing, E.Y., S.-L.C. and L.K.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Telekom Malaysia, under the TM R&D Grant Scheme (No. RDTC/221052).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Restrictions apply to the availability of these data. Data were obtained from Ahmet Emre Aladağ and are available from emre.aladag@boun.edu.tr with the permission of Ahmet Emre Aladağ.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript.

CNN	Convolutional Neural Network
DL	Deep Learning
EIL	Emotion Intensity Lexicon
HTML	Hypertext Markup Language
LIWC	Linguistic Inquiry and Word Count
LR	Logistic Regression
LSTM	Long Short-Term Memory
ML	Machine Learning
NB	Naïve Bayes
NRC	National Research Council Canada
RF	Random Forest

SVM	Support Vector Machines
TF-IDF	Term Frequency-Inverse Document Frequency
UMD	University of Maryland
URL	Uniform Resource Locator
VAD	Valence Arousal Dominance
WHO	World Health Organization
XGBoost	Extreme Gradient Boosting

References

1. World Health Organization. *Suicide Worldwide in 2019: Global Health Estimates*; World Health Organization: Geneva, Switzerland, 2021.
2. World Health Organization. *Preventing Suicide: A Global Imperative*; World Health Organization: Geneva, Switzerland, 2014.
3. O'Connor, R.C.; Nock, M.K. The Psychology of Suicidal Behaviour. *Lancet Psychiatry* **2014**, *1*, 73–85. [CrossRef] [PubMed]
4. American Foundation for Suicide Prevention. *Risk Factors, Protective Factors, and Warning Signs*; American Foundation for Suicide Prevention: New York, NY, USA, 2024; Available online: https://afsp.org/risk-factors-protective-factors-and-warning-signs/ (accessed on 21 July 2022).
5. 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*, 205. [CrossRef] [PubMed]
6. Harmer, B.; Lee, S.; Duong, T.v.H.; Saadabadi, A. Suicidal Ideation. In *StatPearls*; StatPearls Publishing: Treasure Island, FL, USA, 2022.
7. Simon, R.I. Passive Suicidal Ideation: Still a High-Risk Clinical Scenario. *Curr. Psychiatry* **2014**, *13*, 13–15.
8. World Health Organization. *LIVE LIFE: An Implementation Guide for Suicide Prevention in Countries*; World Health Organization: Geneva, Switzerland, 2021.
9. 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.* **2021**, *8*, 214–226. [CrossRef]
10. Gaur, M.; Aribandi, V.; Alambo, A.; Kursuncu, U.; Thirunarayan, K.; Beich, J.; Pathak, J.; Sheth, A. Characterization of Time-Variant and Time-Invariant Assessment of Suicidality on Reddit Using C-SSRS. *PLoS ONE* **2021**, *16*, e0250448. [CrossRef] [PubMed]
11. Alambo, A.; Gaur, M.; Lokala, U.; Kursuncu, U.; Thirunarayan, K.; Gyrard, A.; Sheth, A.; Welton, R.S.; Pathak, J. Question Answering for Suicide Risk Assessment Using Reddit. In Proceedings of the 2019 IEEE 13th International Conference on Semantic Computing (ICSC), Newport Beach, CA, USA, 30 January–1 February 2019; pp. 468–473. [CrossRef]
12. Ji, S.; Yu, C.P.; Fung, S.; Pan, S.; Long, G. Supervised Learning for Suicidal Ideation Detection in Online User Content. *Complexity* **2018**, *2018*, 6157249. [CrossRef]
13. Gaur, M.; Alambo, A.; Sain, J.P.; Kursuncu, U.; Thirunarayan, K.; Kavuluru, R.; Sheth, A.; Welton, R.; Pathak, J. Knowledge-Aware Assessment of Severity of Suicide Risk for Early Intervention. In Proceedings of the World Wide Web Conference—WWW '19, San Francisco, CA, USA, 13–17 May 2019; pp. 514–525. [CrossRef]
14. Renjith, S.; Abraham, A.; Jyothi, S.B.; Chandran, L.; Thomson, J. An Ensemble Deep Learning Technique for Detecting Suicidal Ideation from Posts in Social Media Platforms. *J. King Saud Univ.-Comput. Inf. Sci.* **2021**, *34*, S1319157821003244. [CrossRef]
15. Tadesse, M.M.; Lin, H.; Xu, B.; Yang, L. Detection of Suicide Ideation in Social Media Forums Using Deep Learning. *Algorithms* **2019**, *13*, 7. [CrossRef]
16. Nikhileswar, K.; Vishal, D.; Sphoorthi, L.; Fathimabi, S. Suicide Ideation Detection in Social Media Forums. In Proceedings of the 2021 2nd International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, 7–9 October 2021; pp. 1741–1747. [CrossRef]
17. Li, R.; Deng, R.; Song, T.; Xiao, Y.; Wang, Q.; Fang, Z.; Huang, Y.; Jiao, F.; Chen, Y. Gender-Specific Factors of Suicidal Ideation among High School Students in Yunnan Province, China: A Machine Learning Approach. *J. Affect. Disord.* **2024**, *364*, 157–166. [CrossRef] [PubMed]
18. McHugh, C.M.; Ho, N.; Iorfino, F.; Crouse, J.J.; Nichles, A.; Zmicerevska, N.; Scott, E.; Glozier, N.; Hickie, I.B. Predictive Modelling of Deliberate Self-Harm and Suicide Attempts in Young People Accessing Primary Care: A Machine Learning Analysis of a Longitudinal Study. *Soc. Psychiatry Psychiatr. Epidemiol.* **2023**, *58*, 893–905. [CrossRef] [PubMed]
19. Aladağ, A.E.; Muderrisoglu, S.; Akbas, N.B.; Zahmacioglu, O.; Bingol, H.O. Detecting Suicidal Ideation on Forums: Proof-of-Concept Study. *J. Med. Internet Res.* **2018**, *20*, e215. [CrossRef] [PubMed]
20. Bitew, S.K.; Bekoulis, G.; Deleu, J.; Sterckx, L.; Zaporojets, K.; Demeester, T.; Develder, C. Predicting Suicide Risk from Online Postings in Reddit The UGent-IDLab Submission to the CLPysch 2019 Shared Task, A. In Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology, Minneapolis, MN, USA, 6 June 2019; pp. 158–161. [CrossRef]
21. Shing, H.-C.; Nair, S.; Zirikly, A.; Friedenberg, M.; Daumé, H., III; Resnik, P. 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, New Orleans, LA, USA, 5 June 2018; pp. 25–36. [CrossRef]
22. Ríssola, E.; Ramírez-Cifuentes, D.; Freire, A.; Crestani, F. Suicide Risk Assessment on Social Media: USI-UPF at the CLPsych 2019 Shared Task. In Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology, Minneapolis, MN, USA, 6 June 2019; pp. 167–171. [CrossRef]

23. Shah, F.M.; Haque, F.; Un Nur, R.; Al Jahan, S.; Mamud, Z. A Hybridized Feature Extraction Approach to Suicidal Ideation Detection from Social Media Post. In Proceedings of the 2020 IEEE Region 10 Symposium (TENSYMP), Dhaka, Bangladesh, 5–7 June 2020; pp. 985–988. [CrossRef]
24. Yao, H.; Rashidian, S.; Dong, X.; Duanmu, H.; Rosenthal, R.N.; Wang, F. Detection of Suicidality Among Opioid Users on Reddit: Machine Learning-Based Approach. *J. Med. Internet Res.* **2020**, *22*, e15293. [CrossRef] [PubMed]
25. Aldhyani, T.H.H.; Alsubari, S.N.; Alshebami, A.S.; Alkahtani, H.; Ahmed, Z.A.T. Detecting and Analyzing Suicidal Ideation on Social Media Using Deep Learning and Machine Learning Models. *Int. J. Environ. Res. Public Health* **2022**, *19*, 12635. [CrossRef] [PubMed]
26. Ambalavanan, A.K.; Jagtap, P.D.; Adhya, S.; Devarakonda, M. Using Contextual Representations for Suicide Risk Assessment from Internet Forums. In Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology, Minneapolis, MN, USA, 6 June 2019; pp. 172–176. [CrossRef]
27. Matero, M.; Idnani, A.; Son, Y.; Giorgi, S.; Vu, H.; Zamani, M.; Limbachiya, P.; Guntuku, S.C.; Schwartz, H.A. Suicide Risk Assessment with Multi-Level Dual-Context Language and BERT. In Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology, Minneapolis, MN, USA, 6 June 2019; pp. 39–44. [CrossRef]
28. Haque, F.; Nur, R.U.; Jahan, S.A.; Mahmud, Z.; Shah, F.M. A Transformer Based Approach to Detect Suicidal Ideation Using Pre-Trained Language Models. In Proceedings of the 2020 23rd International Conference on Computer and Information Technology (ICCIT), Dhaka, Bangladesh, 19–21 December 2020; pp. 1–5. [CrossRef]
29. van Kooten, P. Contractions. 2022. Available online: https://github.com/kootenpv/contractions (accessed on 15 March 2023).
30. Liu, M. Are You Really Smiling? Display Rules for Emojis and the Relationship between Emotion Management and Psychological Well-Being. *Front. Psychol.* **2023**, *14*, 1035742. [CrossRef] [PubMed]
31. Montani, I.; Honnibal, M.; Honnibal, M.; Boyd, A.; Van Landeghem, S.; Peters, H. spaCy: Industrial-Strength Natural Language Processing in Python. 2023. Available online: https://github.com/explosion/spaCy (accessed on 15 March 2023).
32. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
33. Tausczik, Y.R.; Pennebaker, J.W. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *J. Lang. Soc. Psychol.* **2010**, *29*, 24–54. [CrossRef]
34. Boyd, R.L.; Ashokkumar, A.; Seraj, S.; Pennebaker, J.W. *The Development and Psychometric Properties of LIWC-22*; University of Texas at Austin: Austin, TX, USA, 2022; pp. 1–47.
35. Mohammad, S.M. Word Affect Intensities. In Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC-2018), Miyazaki, Japan, 7–12 May 2018.
36. Mohammad, S.M. Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words. In Proceedings of the Annual Conference of the Association for Computational Linguistics (ACL), Melbourne, Australia, 15–20 July 2018.
37. Plutchik, R. The Nature of Emotions: Human Emotions Have Deep Evolutionary Roots, a Fact That May Explain Their Complexity and Provide Tools for Clinical Practice. *Am. Sci.* **2001**, *89*, 344–350. [CrossRef]
38. Vishnubhotla, K.; Mohammad, S.M. Tweet Emotion Dynamics: Emotion Word Usage in Tweets from US and Canada. In Proceedings of the Thirteenth International Conference on Language Resources and Evaluation (LREC 2022), Marseille, France, 20–25 June 2022.
39. Kursa, M.B.; Jankowski, A.; Rudnicki, W.R. Boruta—A System for Feature Selection. *Fundam. Informaticae* **2010**, *101*, 271–285. [CrossRef]
40. Homola, D. BorutaPy. 2019. Available online: https://github.com/scikit-learn-contrib/boruta_py (accessed on 6 April 2023).
41. Al-Mosaiwi, M.; Johnstone, T. In an Absolute State: Elevated Use of Absolutist Words Is a Marker Specific to Anxiety, Depression, and Suicidal Ideation. *Clin. Psychol. Sci.* **2018**, *6*, 529–542. [CrossRef] [PubMed]
42. Rnic, K.; Dozois, D.J.A.; Martin, R.A. Cognitive Distortions, Humor Styles, and Depression. *Eur. J. Psychol.* **2016**, *12*, 348–362. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.