Comparative Analysis of Machine Learning and Transformer Models for Detecting Suicidal Ideation in Social Media Texts

Viankail Cedillo-Castelán[1]

[1]National Autonomous University of Mexico, C.U. Coyoacán, 04510, Mexico City, Mexico

Abstract

This study is a contribution of the MentalRiskES 2024 task 3 which explores the detection of suicidal ideation using a range of machine learning models encompassing both traditional algorithms and advanced Transformer models. We assess several approaches including Logistic Regression, Naïve Bayes, Random Forest, and variants of the RoBERTuito model, focusing particularly on the adaptation of the RoBERTuito base model for text analysis and sentiment analysis in Spanish. Results indicate that the RoBERTuito base model outperforms other models in accuracy and performance metrics, demonstrating its potential for critical mental health applications in Spanish-speaking contexts.

Keywords

Suicidal ideation detection, Natural language processing (NLP), Transformer models, RoBERTuito model, Predictive analytics in healthcare

1. Introduction

Mental health and wellness are integral to overall well-being. Mental health research is vital as it influences our emotional, psychological, and social functioning. Mental illnesses, often perceived as personal and isolating struggles, are also significant public health concerns. Society must treat mental disorders with the same urgency and care as chronic medical conditions, recognizing their high treatability and the potential for recovery in many individuals [1]. A better understanding of mental health disorders could lead to improved treatments, interventions, and public health strategies, enhancing mental wellness. Maintaining good mental health is crucial for managing stress, relating to others, and making decisions. With rising global concerns about mental health, there is an urgent need for advanced methods to detect severe mental health conditions early and intervene appropriately. This is particularly crucial for conditions like suicidal ideation—a severe and often hidden symptom associated with many mental health disorders.

Suicidal ideation involves thoughts of self-harm or ending one's life and is a critical mental health challenge with potentially fatal outcomes if unaddressed. This issue affects a diverse range of individuals and is often linked to mental health disorders, though it can also arise from situational stressors. Early recognition of suicidal ideation and understanding its root causes are essential for improving patient outcomes and reducing suicide risks [2]. The World Health Organization (WHO) reports that over 700,000 individuals die by suicide each year, with about one completed suicide for every twenty attempts and many more experiencing suicidal thoughts [3].

The expression and detection of suicidal intent can be facilitated through the analysis of social media posts using natural language processing (NLP) and machine learning (ML) algorithms [4, 5]. Progress in this field, as shown by researchers like Reece & Danforth [6], has been significant. Yet despite advancements in NLP and ML that enable the analysis of textual content for signs of mental health concerns, much of the research, including studies by Coppersmith et al. [7], has focused on specific demographic data. Mental health is deeply influenced by biological, economic, social, and ethnic factors.

IberLEF 2024, September 2024, Valladolid, Spain

viancastelan.edu@gmail.com (V. Cedillo-Castelán)
0000-0002-8287-0469 (V. Cedillo-Castelán)

While some studies have begun to analyze mental health based on gender and social status, attention to different ethnicities remains limited. Notably, research on suicidal ideation among Hispanic populations is scarce.

This study aims to address this gap by evaluating various ML models in NLP, including traditional classifiers like Logistic Regression, Naïve Bayes, and Random Forest, as well as advanced Transformer-based models such as RoBERTuito, which are specially designed to understand the nuances of mental health discourse in Spanish. The effectiveness of RoBERTuito in detecting suicidal thoughts in social media texts underscores its potential in public health strategies for suicide prevention, highlighting the need for culturally relevant mental health interventions within Spanish-speaking communities.

2. Methods

2.1. Data Collection and Usage

For the development of this analysis, the task organizers provided only the test dataset for Task 3; the same procedure was followed in previous corpus [8]. Consequently, to train and validate the models, it was necessary to compile a training dataset from external sources. The primary criteria for data selection were the relevance to the task, the quality and reliability of the data, and the ease of access and use in compliance with legal and ethical standards.

Given these requirements, the training dataset for Task 1 (Disorder detection) for the MentalRiskES [8] at IberLEF 2024 [9] consisting of 727 social media messages in Spanish language was used along with other complementing datasets from published literature. Articles were selected based on their content relevance to the task's domain, which involves identification of suicidal ideation on social media posts. This approach was guided by the need to train our models on data that closely mirrors the characteristics of the test set provided by the task organizers.

Publicly available data was taken from the published tables from studies on the recognition of suicidal intent among depressed populations consisting of 20 social media messages in English language [10] and the screening for suicide risk using social media content consisting of 19 social media messages in Spanish language [7].

To adapt the datasets found in the previously mentioned literature, social media content was translated from its original version in English to Spanish using neural machine translation tools, specifically Google Translate and DeepL. These tools were selected based on their proven effectiveness and accessibility as documented in previous studies [11, 12]. The use of these freely available online services allowed for a consistent and efficient translation process, ensuring that the context captured in the original text was preserved. This approach aligns with methodologies employed in prior research [11, 12].

Model training and evaluation involved splitting the data into training and testing sets using `train_test_split`. The model was trained on a sub-portion (80%) of the collected training set and evaluated on another sub-portion (20%) of the collected training set that was called the testing set.

2.2. Computational Environment

The analysis was conducted using Python 3.9.6. Below are the primary libraries that were utilized for traditional classifiers such as logistic regression, random forest, Naïve Bayes, and SVC models:

- **General Libraries:** Pandas, NumPy, scikit-learn (version 0.24.1), Nltk, Genism, SciPy (version 1.6.0), os, re, json.
- **Specific Tools for Text Feature Extraction:** TfidfVectorizer, CountVectorizer, Word2Vec (via gensim).
- **Miscellaneous:** TensorFlow, Keras, transformers, datasets, accelerate.

2.3. Model Training and Evaluation

For the logistic regression model, text data were preprocessed using TfidfVectorizer and CountVectorizer. For the implementation of the Naive Bayes classification model, the Multinomial Naive Bayes algorithm provided by the scikit-learn library in Python was utilized. For the random forest model, text data were converted into numerical vectors using the word2vec-google-news-300 model from gensim, focusing on semantic content. Stopwords were removed from the text data in the random forest model using the nltk library, which provides a comprehensive list of stopwords for various languages. For the LSTM classifier, the TensorFlow and Keras frameworks were employed to construct and train the model.

To address the complexities of language and sentiment analysis inherent in this task, a transition to utilizing transformer-based models was done.

For the classification tasks involving RoBERTa (Robustly Optimized BERT Pretraining Approach), we utilized the transformers library. RoBERTa, an optimized version of BERT, enhances pretraining techniques to achieve more robust performance across a wider range of tasks. The components utilized in this model setup included:

- **RoBERTa:** RobertaTokenizer, TFRobertaForSequenceClassification.
- **RoBERTuito sentiment analysis and RoBERTuito base pretrained models:** AutoModelForSequenceClassification, AutoTokenizer, Trainer, TrainingArguments.

2.3.1. Model Implementation

Multinomial Naive Bayes: This model was used for the classification with discrete features (word counts for text classification). We trained this model with and without smoothing (alpha parameter). The effectiveness of the model was evaluated using accuracy, precision, recall, and f1-score for the standard version and area under the curve (AUC) for the version with smoothing.

Logistic Regression, Random Forest, and SVC: These models were applied after transforming the trained dataset by TfidfVectorizer. For the SVC model, the impact of document length as an additional feature was explored by combining it with the tf-idf matrix using hstack. Both models were evaluated based on the AUC metric, which provides an aggregate measure of performance across all classification thresholds. Additional features like document length and digit count to the feature set for the Logistic Regression model were implemented to observe any potential improvement in model performance. Finally, for a deeper linguistic analysis, pre-trained word embeddings from the word2vec-google-news-300 model were used to convert messages into average vector representations. Both Logistic Regression and Random Forest classifiers were trained on these embeddings.

LSTM: TensorFlow and Keras frameworks were employed due to their advanced support for deep learning applications. The model architecture was built using the Sequential API, which included several layers crucial for text processing and sequential learning:

- An Embedding layer to transform text inputs into dense vectors of fixed size, capturing semantic information.
- An LSTM layer to learn long-term dependencies within the text data.
- A SpatialDropout1D layer to prevent overfitting by dropping entire 1D feature maps during training.
- A Dense layer to produce the probability distribution over the target classes.

Preprocessing of text data involved tokenizing the texts and padding them to a uniform length to ensure consistent input dimensions across all samples. This was facilitated by Keras' text preprocessing utilities. Additionally, categorical target variables were encoded using scikit-learn's LabelEncoder, adapting them for model training. The model's performance was evaluated using standard metrics from scikit-learn, including accuracy score and a detailed classification report, providing insights into the precision, recall, and F1-scores across different classes.

RoBERTa: The RoBERTa pretrained model on the collected dataset for training aimed at identifying indications of suicidal ideation in social media text achieved an overall accuracy of 66%. The model exhibited a macro precision of 66.07% and a macro recall of 65.85%, resulting in a macro F1-score of 65.81%. The micro-averaged precision, recall, and F1-score all stand at 66%.

RoBERTuito sentiment analysis: The fine-tuning of the RoBERTuito sentiment analysis model on the collected dataset for training aimed at identifying indications of suicidal ideation in social media text achieved an overall accuracy of 74.65%. The model exhibited a macro precision of 74.65% and a macro recall of 75.14%, resulting in a macro F1-score of 74.59%. The micro-averaged precision, recall, and F1-score all stand at 74.65%.

RoBERTuito base uncased: The fine-tuning of the RoBERTuito base uncased model on the collected dataset for training aimed at identifying indications of suicidal ideation in social media text achieved an overall accuracy of 89.61%. The model exhibited a macro precision of 94.59% and a macro recall of 63.64%, resulting in a macro F1-score of 68.57%. The micro-averaged precision, recall, and F1-score all stand at 89.61%.

3. Results

Traditional classifiers: Multinomial Naive Bayes, Logistic Regression, Random Forest, and SVC demonstrated suboptimal performance. Specifically, the Multinomial Naïve Bayes model achieved a modest accuracy of 57.93%, and the SVC model slightly better at 59.8%. The Logistic Regression model performed somewhat more effectively, reaching an accuracy of 63%. Notably, the Random Forest model outperformed the others with an accuracy of 69%.

LSTM model on the collected dataset for training aimed at identifying suicidal ideation in social media text achieved an overall accuracy of 61.6%. The model exhibited a macro precision of 61.6% and a macro recall of 61.1%, resulting in a macro F1-score of 60.9%. The micro-averaged precision, recall, and F1-score all stand at 61%.

RoBERTa pretrained model on the collected dataset for training aimed at identifying indications of suicidal ideation in social media text achieved an overall accuracy of 66%. The model exhibited a macro precision of 66.07% and a macro recall of 65.85%, resulting in a macro F1-score of 65.81%. The micro-averaged precision, recall, and F1-score all stand at 66%.

RoBERTuito sentiment analysis fine-tuned model on the collected dataset for training aimed at identifying indications of suicidal ideation in social media text achieved an overall accuracy of 74.65%. The model exhibited a macro precision of 74.65% and a macro recall of 75.14%, resulting in a macro F1-score of 74.59%. The micro-averaged precision, recall, and F1-score all stand at 74.65%.

RoBERTuito base uncased fine-tuned model on the collected dataset for training aimed at identifying indications of suicidal ideation in social media text achieved an overall accuracy of 89.61%. The model exhibited a macro precision of 94.59% and a macro recall of 63.64%, resulting in a macro F1-score of 68.57%. The micro-averaged precision, recall, and F1-score all stand at 89.61%.

4. Discussion

The effectiveness of various classifiers was evaluated in this study, including both traditional models and advanced deep learning approaches, in detecting indications of suicidal ideation within social media posts. Our analysis highlights the varying degrees of success and points to potential areas for further optimization.

Traditional classifiers, including Multinomial Naïve Bayes, SVC, Logistic Regression, and Random Forest, and even LSTM, demonstrated a range of effectiveness, with accuracies ranging from 57.93% to 69%. These models, while historically robust across various text classification tasks, fell short in this context, particularly in handling the complexities of mental health data. The highest performing among them, the Random Forest model, achieved an accuracy of 69%; however, this performance still highlights significant limitations, underscoring the need to explore more sophisticated deep learning models capable of handling the complexities of mental health data, particularly in a delicate topic such as suicidal ideation.

Table 1
Results on the official test dataset provided by MentalRiskES 2024

Model	Accuracy	Macro_P	Macro_R	Macro_F1	Micro_P	Micro_R	Micro_F1
RoBERTuito Base Uncased	0.691	0.345	0.500	0.409	0.691	0.691	0.691

The RoBERTa model achieved an accuracy of 66% on the collected training dataset, which indicates that it correctly predicts the binary labels for only two-thirds of the dataset. While this demonstrates some capability to generalize from training data to unseen data, the accuracy suggests that there might be substantial room for improvement.

The RoBERTuito sentiment analysis fine-tuned model achieved an accuracy of 74.65% on the collected training dataset, indicating a relatively high level of correctness in classifying text as either indicative or non-indicative of suicidal ideation. This suggests that the model is effectively leveraging its sentiment analysis training to make accurate predictions in a new closely related domain. Nevertheless, there is still potential for enhancement.

The election of RoBERTuito base uncased after trying RoBERTuito Sentiment Analysis model was due to the nature of the multiclass classifier of RoBERTuito Sentiment Analysis. Since it has been specifically trained to classify text into multiple sentiment categories such as positive, negative, and neutral, it may have still a bias even after indicating a binary classification.

In contrast, RoBERTuito Base Uncased is not specifically a multiclass classifier but a more general-purpose model pretrained on a language modeling task without specific tuning for sentiment analysis or any other specialized classification task. Here it was taken as a foundational model that was adapted to this specific binary classification.

RoBERTuito Base Uncased resulted in an accuracy of 89.61% on the test dataset of our collected training data, which indicates that the model is effectively generalizing from the training data to the unseen test data, demonstrating robust performance in distinguishing between texts that either suggest or do not suggest suicidal ideation.

The macro-averaged precision of 94.59% indicates that, on average, the model's predictions of each class are reliable. However, the macro-averaged recall of 63.64% suggests that the model, while precise, is conservative in predicting positive instances and could be missing a proportion of actual positive cases. Nevertheless, the similarity between micro-averaged metrics and overall accuracy suggests that the model performs uniformly across all instances, providing confidence in its general utility.

Testing our RoBERTuito Base Uncased fine-tuned model on the testing data provided by MentalRiskES, the performance of the was evaluated across three runs in comparison to other teams and a baseline model. The results demonstrated that our RoBERTuito Base Uncased fine-tuned consistently achieved an accuracy of 0.691 across all runs. In terms of macro-averaged metrics, the RoBERTuito Base Uncased fine-tuned attained a Macro_Precision (Macro_P) of 0.345, Macro_Recall (Macro_R) of 0.500, and Macro_F1 score of 0.409. Similarly, for micro-averaged metrics, the RoBERTuito Base Uncased fine-tuned recorded a Micro_Precision (Micro_P), Micro_Recall (Micro_R), and Micro_F1 score all at 0.691. The RoBERTuito Base Uncased fine-tuned performance on the Early Risk Detection Error (ERDE) metrics showed an ERDE5 score of 0.261 and an ERDE50 score of 0.261. Additionally, the RoBERTuito Base Uncased fine-tuned latency to positive prediction (latencyTP) was consistent at 0.214 with a speed of 1 and a latency-weighted F1 score of 0.817. See Table 1. Overall, our RoBERTuito Base Uncased fine-tuned performance was comparable across all runs, maintaining a balance between precision and recall with strong latency-weighted F1 scores, highlighting the consistency and reliability of their approach in this evaluation.

While traditional classifiers such as Logistic Regression, Naïve Bayes, and Random Forest have demonstrated utility in handling a variety of text classification tasks, the evolution of deep learning has introduced more sophisticated models that can capture intricate patterns in large datasets. This shift marks a significant advancement in the field of natural language processing.

5. Conclusion and Future Directions

While the transition from traditional classifiers to sophisticated deep learning models like RoBERTuito represents a significant leap forward in our capacity to process and analyze mental health data, future efforts need to continue. Particularly in Hispanic-speaking communities where little research has been performed; the challenge of accurately detecting suicidal ideation cannot be focused on specific demographic groups. The results of this study underscore the pressing need for continued enhancements, especially in improving recall without sacrificing precision, to effectively support suicide prevention efforts and provide timely interventions in Hispanic-speaking populations where such resources are critically needed.

6. Limitations

When testing the RoBERTuito base uncased fine-tuned model on the actual test dataset provided by the MentalRiskES committee at IberLEF 2024, the metrics resulted in an accuracy of 69.1%, a macro-averaged precision of 34.5%, a macro-averaged recall of 50%, and a macro-averaged F1 score of 40.9%; and a micro-averaged precision of 69.1%, a micro-averaged recall of 69.1%, and a micro-averaged F1 score of 69.1%. This could be due to the differences between the training and the test datasets. This approach was guided by the need to train our models on data that closely mirrors the characteristics of the test set provided by the task organizers, albeit the exact content and context might differ.

Acknowledgments

This research was supported by the National Autonomous University of Mexico. Special thanks to the MentalRiskES 2024 task organizers and the IberLEF 2024 committee.

References

[1] S. K. Galson, Mental health matters, Public Health Reports 124 (2009) 189–191.

[2] B. Harmer, S. Lee, T. V. H. Duong, A. Saadabadi, Suicidal ideation, StatPearls (2020).

[3] W. H. Organization, Suicide, Fact Sheets (2023).

[4] J. Jashinsky, S. H. Burton, C. L. Hanson, J. West, C. Giraud-Carrier, M. D. Barnes, T. Argyle, Tracking suicide risk factors through twitter in the us, Crisis (2014).

[5] T. H. Aldhyani, S. N. Alsubari, A. S. Alshebami, H. Alkahtani, Z. A. Ahmed, Detecting and analyzing suicidal ideation on social media using deep learning and machine learning models, International journal of environmental research and public health 19 (2022) 12635.

[6] A. G. Reece, C. M. Danforth, Instagram photos reveal predictive markers of depression, EPJ Data Science 6 (2017) 15.

[7] G. Coppersmith, R. Leary, P. Crutchley, A. Fine, Natural language processing of social media as screening for suicide risk, Biomedical informatics insights 10 (2018) 1178222618792860.

[8] A. M. Mármol-Romero, A. Moreno-Muñoz, F. M. P. del Arco, M. D. M. González, M. T. M. Valdivia, L. A. Ureña-López, A. M. Ráez, Overview of mentalriskes at iberlef 2024: Early detection of mental disorders risk in spanish, in: Procesamiento del Lenguaje Natural, volume 73, 2024.

[9] R. F. Chiruzzo L., Jiménez-Zafra S. M., Proceedings of the iberian languages evaluation forum (iberlef 2024) co-located with the 40th conference of the spanish society for natural language processing (sepln 2024), CEUR-WS.org, 2024.

[10] S. B. Hassan, S. B. Hassan, U. Zakia, Recognizing suicidal intent in depressed population using nlp: a pilot study, in: 2020 11th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), IEEE, 2020, pp. 0121–0128.

[11] E. Steigerwald, V. Ramírez-Castañeda, D. Y. Brandt, A. Báldi, J. T. Shapiro, L. Bowker, R. D. Tarvin, Overcoming language barriers in academia: Machine translation tools and a vision for a multilingual future, BioScience 72 (2022) 988–998.

[12] L. Bowker, Promoting linguistic diversity and inclusion, The International Journal of Information, Diversity, Inclusion 5 (2021) 127–151.

A. Online Resources

The sources for the ceur-art style are available via:

- GitHub
- Overleaf template