A comparative analysis of transfer learning models on suicide and non-suicide textual data

Merinda Lestandy[1], Abdurrahim[2], Amrul Faruq[1], Muhammad Irfan[1]
[1]Department of Electrical Engineering, Faculty of Engineering, Universitas Muhammadiyah Malang, Malang, Indonesia
[2]Informatics Master Program, Faculty of Industrial Technology, Islamic University of Indonesia, Daerah Istimewa Yogyakarta, Indonesia

Article Info

Article history:

Received Aug 27, 2024
Revised Jan 6, 2025
Accepted Jan 22, 2025

Keywords:

Deep learning
Media social
Sentiment analysis
Suicide
Transformer model

ABSTRACT

The rise of social media has allowed individuals to express themselves freely, increasing the visibility of mental health concerns, including suicidal tendencies. This issue is particularly significant, as suicide is one of the leading causes of death globally. The objective of this study is to develop a model capable of accurately detecting suicide-related textual data using advanced natural language processing techniques. To achieve this, we applied transfer learning models, including bidirectional encoder representations from transformers (BERT), robustly optimized bidirectional encoder representations from transformers (RoBERT), a lite BERT (ALBERT), and decoding-enhanced BERT with disentangled attention (DeBERTa). the dataset used in this research includes 232,074 posts from Reddit, categorized into suicide and non-suicide labels. Preprocessing steps such as removing HTML tags, special characters, and punctuation were applied, followed by stopword removal and lemmatization. The models were trained and evaluated using accuracy, precision, recall, and F1-score metrics. Among the models tested, DeBERTa demonstrated superior performance, achieving an accuracy of 98.70% and an F1-score of 98.70%. These findings suggest that transfer learning models, particularly DeBERTa, are effective in identifying suicidal ideation in textual data.

Corresponding Author:

Merinda Lestandy
Department of Electrical Engineering, Faculty of Engineering, Universitas Muhammadiyah Malang
Malang, Indonesia
Email: merindalestandy@umm.ac.id

1. INTRODUCTION

The issue of mental health has a global impact, exerting substantial effects on countries throughout the spectrum of development. According to the mental health action plan for the years 2013-2020, as reported by the World Health Organization (WHO), approximately one out of every four individuals globally encounters mental diseases of various severity [1]. Unfortunately, a significant proportion of individuals experiencing mental health difficulties, namely 75%, do not obtain sufficient therapy, hence increasing their existing challenges. In the past few years, there has been a growing recognition of the heightened importance of emotional concerns associated with coronavirus disease 2019 (COVID-19) and the experience of isolation [2]. This is particularly noteworthy in the context of individuals who already have pre-existing mental health illnesses [3]. The aforementioned phenomenon has resulted in elevated levels of anxiety, despair, substance misuse, social isolation, intimate partner violence, and, in severe instances, self-inflicted mortality [4]. It is worth mentioning that there has been an increase in the probability of attempted suicide within the general population [2].

Based on data from the WHO, it has been determined that around 800,000 individuals engage in acts of suicide annually, with the age cohort ranging from 15 to 29 years exhibiting the highest incidence rates [5]. This exemplifies the fact that suicide ranks as the second most prominent cause of mortality among adolescents on a global scale. Currently, social media platforms enable interactive communication, enabling users to share their views and emotions through the means of posts and comments. Social media platforms can function as a medium for disseminating information regarding risk factors associated with suicide. The analysis includes environmental factors, such as instances of abuse and exposure to stressful events; health-related challenges, such as the presence of chronic diseases and mental health disorders; and historical factors, including family history and previous suicide attempts. All these qualities are positively associated with suicidal intent. Moreover, there exist various factors, specifically three primary risk factors, that can contribute to the development of suicidal intent or ideation: environmental factors (such as abuse and highly stressful life events), health factors (including chronic pain and mental health issues), and historical factors (such as a family history of suicide or previous suicide attempts). These risk factors can potentially lead to the development of suicidal intentions or thoughts [6]. Therefore, the field of natural language processing (NLP) assumes a significant role in the detection and identification of suicide intentions expressed in textual content. The provided data presents valuable insights that can be utilised in the development of systems that have the capability to forecast an individual’s propensity for engaging in suicide attempts [7]. Researchers utilise many types of models, including machine learning models, deep learning models, and transformer-based models, in order to identify textual content that indicates potential suicide thoughts [6]. These efforts contribute to providing protection and reducing the incidence of suicide.

In recent years, there has been a significant focus in the field of NLP on the automated identification of mental health disorders using diverse data sources such as electronic health records (EHRs), clinical records, biomarkers, written texts, and online posts [8]-[10]. NLP methodologies have exhibited their efficacy in the classification of initial indications of mental illness, akin to the study of sentiment [11]-[13]. The transformer-based language model [14] is an approach that has demonstrated remarkable effectiveness. The transduction model under discussion is founded upon a phenomenon known as attention, which has significantly transformed brain encoders designed for natural language sequences. The transformer architecture omits any noise or convolutional structures, hence allowing the acquisition of sequence information in the input exclusively through attention mechanisms. The presence of a self-attention mechanism within the encoder’s processing block is responsible for this outcome. The bidirectional context information processing issue, which involves processing word and sequence inputs simultaneously, can be effectively addressed by transformers [14]-[16]. Sophisticated contextual language understanding models can be acquired, which could catch subtle and detailed lexical patterns. This results in the development of a comprehensive feature representation of a given text [17]. Several scholarly research [18]-[21] have investigated the application of transformers in the field of NLP, specifically in the domain of suicide-related text identification [6], [22], [23]. Hence, this research makes svaluable contribution to the subsequent domains, includings analyze the suicidal ideation text datasets, it is recommended to utilize advanced deep learning models like bidirectional long short-term memory (BiLSTM), as well as various transfer learning models such as bidirectional encoder representations from transformers (BERT), robustly optimized BERT approach (RoBERTa), a lite BERT (ALBERT), decoding-enhanced BERT with disentangled attention (DeBERTa). This strategy seeks to investigate the efficacy of these models in identifying suicidal ideation. Furthermore, this research enhances to create more sophisticated transfer learning models in contrast to conventional deep learning models. This breakthrough incorporates pre-training and fine-tuning methodologies that greatly enhance the ability to detect suicidal thoughts in written language. In addition, this study also contributes that DeBERTa exhibits robust performance and is on par with comparable models in terms of competitiveness.

2. RELATED WORKS

Recent studies have investigated the correlation between mental well-being and the way people express themselves linguistically to detect signs of suicidal ideation. Prior research incorporated language components derived from psychiatric literature, including linguistic inquiry and word count (LIWC) [24], emotional characteristics [25], and suicide letters [26]. Nevertheless, these methodologies are constrained in their ability to assess individual posts and are inadequate for datasets that are diverse or extensive [27]. The utilization of social media and NLP in the field of mental health research has experienced a notable surge in popularity. The study of sentiment analysis is expanding to include online mental health forums and social media data as new areas of investigation. Tadesse *et al.* [28] devised a hybrid model that integrates latent dirichlet allocation (LDA), linguistic inquiry and word count analysis (LIWCA), Bigram, and multilayer perceptron (MLP), achieving a commendable accuracy rate of 90%. Additional research [29]-[31] gathered information from Twitter and using different machine learning techniques to categorize suicidal ideation. Deep learning techniques such as long short-term memory (LSTM) and convolutional neural network (CNN) have made substantial progress in NLP because of the widespread use of word embeddings. ML approaches are constrained by restrictions such as the curse of dimensionality, data sparsity, and time consumption, which render them inappropriate for some applications. Deep learning improves upon conventional machine learning methods by extracting higher-level features from input data using a greater number of layers in the model, resulting in more reliable and precise classification. Studies indicate that deep learning models have superior prediction accuracy in identifying suicidal ideation when compared to other machine learning classifiers [32], [33].

Presently, numerous researchers are utilizing BERT-based models [23] to identify suicidal ideation [34], as these models have the ability to precisely capture semantic and contextual characteristics [35]. Nevertheless, suicide detection research is constrained by its emphasis on Twitter, a platform that restricts writing to a maximum of 280 characters. Previous research employing transformer models [6] employed suicide-related data from Twitter as the dataset. The models utilized in these investigations were BERT, DistilBERT, ALBERT, RoBERTa, and DistilRoBERTa. Among these models, RoBERTa demonstrated the highest level of accuracy, attaining a score of 95.39%. A different study [36] employed a merged RoBERTa and CNN model, with a dataset consisting of 110,040 data points. The composite model attained an F1-score of 96.81%. This work tries to create a modified version of the BERT model that addresses the limits of huge datasets and focuses on combining several models rather than modifying the BERT model itself. This study aims to construct models using various BERT variations, including BERT, RoBERTa, ALBERT, and DeBERTa. Notably, the DeBERTa model has not been examined in previous studies.

3. METHOD

The investigation commenced through several methodological stages, as illustrated in Figure 1. The dataset included of texts pertaining to suicide and non-suicide. The initial step involved preprocessing the text by eliminating special characters, numerical values, and punctuation marks, followed by converting it to lowercase. We conducted lemmatization to reduce words to their base form. We utilized the processed data in pre-trained models such as BERT, RoBERTa, ALBERT, and DeBERTa, adapting them to specific analysis tasks via transfer learning. We categorized the outcomes to anticipate data classifications. We assessed the accuracy of the model by utilizing a confusion matrix and performed model comparisons to determine the most ideal performance. This approach integrates text processing methodologies with deep learning algorithms to achieve efficient analysis.

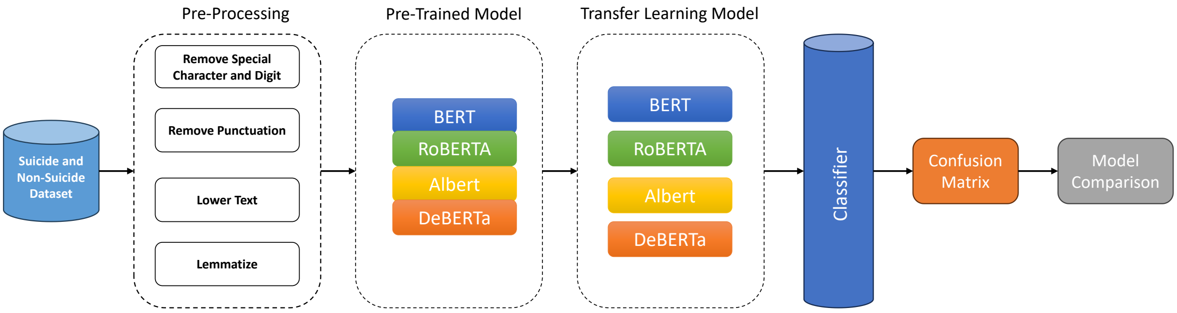


Figure 1. System architecture of the proposed work

3.1. Preprocessing

Preprocessing is a critical stage in ensuring that the textual data is in optimal form for analysis and model training. In this study, we began by removing unnecessary elements such as HTML tags, special characters, numerals, and punctuation marks. This was necessary to eliminate extraneous symbols that do not contribute meaningfully to the linguistic structure, ensuring that the models could focus on essential textual patterns without being distracted by irrelevant noise [37].

After cleaning the dataset of unnecessary characters, the text was standardized by converting all words to lowercase. This step plays a significant role in reducing redundancy caused by case differences, especially for languages like English, where capitalization does not change the meaning of words [38]. This standardization helps in minimizing discrepancies between words like “Suicide” and “suicide,” treating them as identical during model training [39].

Subsequently, stopword removal was conducted. Stopwords are common words, such as “the,” “is,” and “in,” which, while essential for grammatical correctness, do not carry significant meaning for text classification tasks. By removing these stopwords, we ensure that the model concentrates on the more critical and informative words, thus enhancing its performance [40]. Additionally, this step reduces the dimensionality of the dataset, allowing for more efficient computation [41].

The final step in preprocessing involved lemmatization, where words are reduced to their base or root forms. This process ensures that words like "running," "runs," and "ran" are treated as "run," preventing the model from treating them as separate entities. Lemmatization thus aids in capturing the core meaning of words and improves the model's ability to generalize patterns from the data [42].

3.2. Tranformer

BERT, as described in the literature [15], is a notable breakthrough in the field of NLP, owing to its utilization of the transfer learning model. The model consists of a sequence of transformer encoder layers, and it offers several advantages, such as a cohesive architecture for diverse applications and bidirectional pre-training. The primary task of self-supervised pre-training, referred to as masked language modeling (MLM), plays a crucial role in enabling bidirectionality inside the model. It involves predicting masked words in unlabeled text by considering both the preceding and succeeding context words concurrently. In addition to its primary pre-training objective, BERT has also been equipped with the capability to effectively process numerous sequences by means of the next sentence prediction (NSP) work. The NSP task is designed to discern whether a given sequence serves as a continuation of another sequence. Numerous enhancements have been devised after the introduction of BERT. RoBERTa [43] is a refined version of BERT, a popular language model, which incorporates a more efficient training methodology. To improve contextual understanding in the text generated by the model, RoBERTa eliminates the NSP task. Additionally, she enhances the training process by augmenting sentence pairings through random modifications using a large corpus of text data. In addition, RoBERTa employs prolonged training durations and increased batch sizes to improve the performance of the model.

In the year 2019, a transformer-based model called ALBERT [44] was created, featuring a reduction in its parameter count. The objective of this BERT addition is to enhance efficiency while maintaining the quality of linguistic context comprehension. ALBERT employs a parameter sharing technique among layers in the transformer architecture, resulting in a reduction in the number of parameters that need to be trained. By employing this methodology, ALBERT was able to attain similar, if not better, results compared to BERT, all while utilizing a smaller number of parameters. The year 2020 witnessed the emergence of a novel model known as DeBERTa [45], which introduced modifications to the attention mechanism inside the transformer architecture, hence enhancing the capabilities of transformer-based language models. The DeBERTa model employs a disentangled attention mechanism that effectively divides attention into two components: content-based attention and position-based attention. This methodology facilitates the resolution of certain obstacles encountered by transformer-based models, including the reliance on inflexible word sequencing and the incapacity to comprehend nonlinear associations among words. By incorporating the disentangled attention mechanism, DeBERTa demonstrates enhanced text representations and improved performance in many natural languages processing tasks, including text comprehension, sentiment analysis, and classification tasks.

3.3. Evaluation matrix

The evaluation of the model will be conducted using a separate set of testing data, and the results will be given in the form of a confusion matrix. The performance of this algorithm will be evaluated based on metrics such as accuracy, precision, recall, and F-1 score. The measure of accuracy is determined by the degree of proximity between the projected value and the actual value, as shown in (1). The term "true positive" (TP) is used to describe instances when positive data is accurately predicted, while "true negative" (TN) is used to describe instances where negative data is accurately predicted. The term "false positive" (FP) refers to instances where negative data is wrongly classified as positive data, while "false negative" (FN) denotes cases where positive data is incorrectly classified as negative data. In (2) represents precision, which is a metric used to assess the accuracy of data by evaluating the proportion of correct predictions.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{1}$$

Precision is a measure that quantifies the degree of precision exhibited by data, considering both the veracity of the information and its ability to make accurate predictions. In (2) is utilized to articulate the concept of precision.

$$Precision = \frac{TP}{TP+FP} \tag{2}$$

The effectiveness of the model in accurately identifying a certain class can be evaluated by its recall metric, which can be computed using (3).

$$Recall = \frac{TP}{TP+FN} \tag{3}$$

The F1-score is a metric that integrates two fundamental ideas in model evaluation, namely precision and recall. Precision assesses the degree of accuracy in the data projected by the model, whereas recall evaluates the efficacy of the model in correctly recognizing a specific class. The F1-score use (4) to evaluate the overall effectiveness of the model in accurately predicting outcomes.

$$F1 - Score = 2\ \frac{(Recall \times Precision)}{(Recall + Precision)} \tag{4}$$

4. RESULTS AND DISCUSSION

4.1. Dataset

Table 1 shows the dataset that was acquired from the social media network Reddit and has been categorized into two distinct labels: label 1 denotes instances of suicide, while label 0 represents non-suicidal occurrences. The dataset utilized in this study consists of two distinct subreddits. The data collection process involved utilizing the Pushshift API to retrieve posts from the SuicideWatch subreddit, spanning the time from December 16th, 2008, to January 2nd, 2021. The data collection process involved utilizing the Pushshift API to retrieve posts from the SuicideWatch subreddit, encompassing the time spanning from December 16th, 2008, to January 2nd, 2021. A collection of posts pertaining to non-suicidal subjects was compiled from the subreddit r/teenagers. The dataset comprises a total of 232,074 entries, which are categorized into two groups: non-suicide data labels (116,037 entries) and suicide data labels (116,037 entries). The following table provides an illustration of both the datasets pertaining to suicide and non-suicide cases.

Table 1. Example results of non-suicide and suicide datasets

Text	Label
My apologies to the random dude on Among Us So uhh.... I would like to apologize to the random dude named Pon on among us. I had hosted a game for my gf and her friend to learn how to play and whatever and... we may have taken over the chat with UwU s and uh.... other questionable language.... * cough * I was called “mommy yellow” * cough * ANYWAYS sorry ya had to witness that dude.	0
As my isolation grows longer, thoughts of murder and rape enter my mind. The humans I see every day start to look less and less like the same species as me. They start to feel like just another animal. I start to fantasize about how I could corner one of them and attack them. Satisfy some of my carnal needs. I don’t want to be this way, or have these thoughts. But the less real affection I receive, the more attractive and comforting these thoughts start to be. Why would somebody want me alive? I’m like a dangerous animal. I wish somebody would come and kill me. I would be better off dead I’m sure. And anybody I’m around would be better off too. My urges to kill myself have always been stronger than my urges to hurt others. I’ve come close to killing myself. Standing at the edge of tall structures. I go with the intent to jump, and I could if I had just made one sharp and sudden movement. I’ve also thrown myself in traffic, but it was pretty low speed and I got up without a scratch. The closest I’ve got to hurting anybody was when I saw a pretty girl walking at night and I started to follow her on my bike. I could tell she noticed me and was scared. But I feel like it was a far cry from actually interacting with her, let alone hurting her. I assume as I my period of isolation and loneliness continues, I will start to get more courage to do both. But I really hope I kill myself before hurting anybody.	1

4.2. Pre-processing

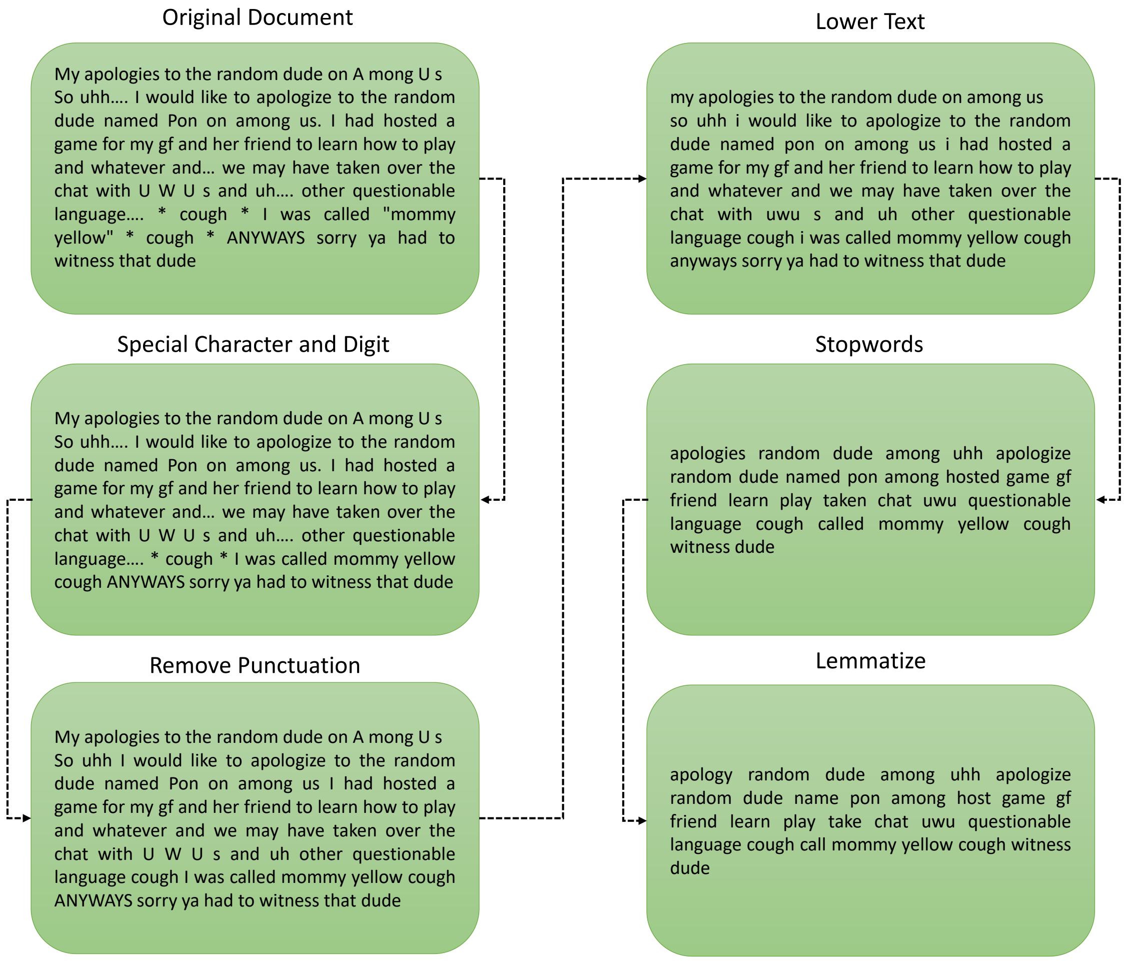
The pre-processing stage occurs after the gathering of data. During this stage, several cleaning procedures are used to the data. These procedures involve the elimination of HTML elements, special characters, numerals, and punctuation. Additionally, the data is converted to lowercase and stopwords are removed. The results of the pre-processing procedure are depicted in Figure 2.

4.3. Test result

The evaluation results of deep learning and transfer learning models are presented in Table 2. We employed the BiLSTM model for deep learning and evaluated the performance of BERT, RoBERTa, ALBERT, and DeBERTa as transfer learning models. The DeBERTa model attained a remarkable accuracy score of 0.9870%, showcasing consistent performance in terms of precision, recall, and F1-score. The subsequent model, RoBERTa, attained an accuracy score of 0.9834%, whilst BERT had a little lower score of 0.9800%. ALBERT attained a precision score of 0.9500%. These findings demonstrate that DeBERTa possesses exceptional skills and can rival other models, as evidenced by the results presented in Table 2 of the testing data.

In contrast to prior research [36], there is a notable enhancement. The prior study documented an F1-score of 96.81%. This enhancement indicates that the fundamental models created in this research are on par with composite models like RoBERTa-CNN. The inclusion of this feature enables the transformer model to surpass the constraints in speed and efficiency that are seen in recurrent models like BiLSTM. The BiLSTM model faces challenges in collecting non-linear correlations between words that are widely separated in the text, hence hindering its ability to accurately express intricate semantic connections. The table’s results demonstrate that the utilization of transfer learning has the capability to overcome the limitations linked to the BiLSTM model.

The objective of this work is to assess the effectiveness of several types of transformer models in text classification tasks that involve sensitive topics, such as identifying indications of suicide on social media. The results underscore the significance of choosing the suitable model to enhance performance in tasks related to NLP, particularly in extremely crucial and influential circumstances such as emotional datasets. Nevertheless, there are still constraints regarding the performance of these models on diverse datasets or in changing circumstances. Subsequent investigations could prioritize additional exploration and more extensive testing to enhance the overall applicability of these models.



Original Document

My apologies to the random dude on A mong U s So uhh…. I would like to apologize to the random dude named Pon on among us. I had hosted a game for my gf and her friend to learn how to play and whatever and… we may have taken over the chat with U W U s and uh…. other questionable language…. * cough * I was called "mommy yellow" * cough * ANYWAYS sorry ya had to witness that dude

Special Character and Digit

My apologies to the random dude on A mong U s So uhh…. I would like to apologize to the random dude named Pon on among us. I had hosted a game for my gf and her friend to learn how to play and whatever and… we may have taken over the chat with U W U s and uh…. other questionable language…. * cough * I was called mommy yellow cough ANYWAYS sorry ya had to witness that dude

Remove Punctuation

My apologies to the random dude on A mong U s So uhh I would like to apologize to the random dude named Pon on among us I had hosted a game for my gf and her friend to learn how to play and whatever and we may have taken over the chat with U W U s and uh other questionable language cough I was called mommy yellow cough ANYWAYS sorry ya had to witness that dude

Lower Text

my apologies to the random dude on among us so uhh i would like to apologize to the random dude named pon on among us i had hosted a game for my gf and her friend to learn how to play and whatever and we may have taken over the chat with uwu s and uh other questionable language cough i was called mommy yellow cough anyways sorry ya had to witness that dude

Stopwords

apologies random dude among uhh apologize random dude named pon among hosted game gf friend learn play taken chat uwu questionable language cough called mommy yellow cough witness dude

Lemmatize

apology random dude among uhh apologize random dude name pon among host game gf friend learn play take chat uwu questionable language cough call mommy yellow cough witness dude

Figure 2. Pre-processed results of suicide and non-suicide datasets

Table 2. Testing results of transformer model

Algoritma model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
BiLSTM+GloVe	0.9346	0.9532	0.9142	0.9332
BERT	0.98	0.98	0.98	0.98
RoBERTa	0.9834	0.9834	0.9834	0.9834
ALBERT	0.95	0.95	0.95	0.95
DeBERTa	0.9870	0.9870	0.9870	0.9870
RoBERTa-CNN [36]	0.98	0.9698	0.964	0.9681

The class division of the confusion matrix outcomes for the DeBERTA model is displayed in Table 3. The matrix encompasses a total of 46,415 testing data instances. The F1-score for the non-suicide label is 0.9871%, indicating that the model's identification of non-suicide text is associated with a minimal number of false positives. In a similar vein, the F1 score for the suicide label wording is 0.9869%. The DeBERTa model provides the advantage of employing disentangled attention, which facilitates a more focused and structured attention mechanism. This approach helps address challenges related to understanding the relationship between words in longer texts and improves the ability to perceive context in a broader range of information. The successful outcome of this has been proven by the findings presented in Table 2 that result from each trial conducted on the transformer models.

Table 3. Class division of the DeBERTa model

Label	Precision	Recall	F1-score
Non-suicide	0.9835	0.9907	0.9871
Suicide	0.9906	0.9832	0.9869

5. CONCLUSION

This study aims to accurately identify suicidal tendencies by textual analysis, which is crucial since suicide is the second most common cause of death among those aged 15 to 29 worldwide, as reported by the WHO. The COVID-19 epidemic has exacerbated mental health difficulties, heightening the likelihood of suicide. Social media has emerged as a crucial platform for those experiencing distress to openly express and disseminate their emotions, thereby serving as a helpful tool for promptly identifying and intervening in such cases. This study utilized deep learning and transfer learning models to accurately classify texts as either suicide-related or non-suicide-related. We implemented meticulous data cleansing procedures, which involved deleting HTML components, special characters, digits, punctuation, and emojis. Additionally, we converted the text to lowercase, removed stopwords, and performed lemmatization on the words. We evaluated our models by comparing them to other sophisticated transformers and deep learning techniques, such as BERT, RoBERTa, ALBERT, and DeBERTa, in addition to the BiLSTM model based on GloVe embeddings. The findings indicated that the DeBERTa model had superior performance compared to other models in terms of accuracy, precision, recall, and F1-score. The reason behind DeBERTa's exceptional performance can be attributed to its disentangled attention mechanism. This technique enables the model to choose to concentrate on pertinent textual components while rejecting extraneous information. This feature improves its efficacy in capturing the intricate and intricate aspects of language associated with suicidal ideation.

The results of our research indicate that advanced transformer models, specifically DeBERTa, show great potential in improving the identification of suicidal tendencies by analyzing text from social media. These findings have significant consequences for early intervention and prevention methods, potentially reducing mortality rates by identifying persons at risk more precisely and swiftly. Subsequent studies should investigate the extent to which these models may be applied to diverse datasets and examine how different data preprocessing approaches affect the performance of the models. Furthermore, incorporating a wide range of linguistic and cultural contexts could significantly strengthen the reliability and practicality of these models in real-life situations.

ACKNOWLEDGEMENTS

This research was supported by the Department of Electrical Engineering, University of Muhammadiyah Malang and the Computer Laboratory of the Department of Electrical Engineering, University of Muhammadiyah Malang, which provided insights and expertise that really helped the research.

REFERENCES

[1] K. Windfuhr and N. Kapur, "Suicide and mental illness: a clinical review of 15 years findings from the UK National Confidential Inquiry into Suicide," *British Medical Bulletin*, vol. 100, pp. 101–121, 2011, doi: 10.1093/bmb/ldr042.

[2] M. A. Reger, I. H. Stanley, and T. E. Joiner, "Suicide Mortality and Coronavirus Disease 2019—A Perfect Storm?," *JAMA Psychiatry*, vol. 77, no. 11, pp. 1093–1094, Nov. 2020, doi: 10.1001/jamapsychiatry.2020.1060.

[3] I. Buneviciene, R. Bunevicius, S. Bagdonas, and A. Bunevicius, "The impact of pre-existing conditions and perceived health status on mental health during the COVID-19 pandemic," *Journal of Public Health*, vol. 44, no. 1, pp. e88–e95, Mar. 2022, doi: 10.1093/pubmed/fdab248.

[4] S. Galea, R. M. Merchant, and N. Lurie, "The Mental Health Consequences of COVID-19 and Physical Distancing: The Need for Prevention and Early Intervention," *JAMA Internal Medicine*, vol. 180, no. 6, pp. 817–818, Jun. 2020, doi: 10.1001/jamainternmed.2020.1562.

[5] World Health Organization, *National suicide prevention strategies.*, vol. 30. 2018. [Online]. Available: https://apps.who.int/iris/bitstream/handle/10665/279765/9789241515016-eng.pdf?ua=1 (Accessed: Jul. 01, 2023)

[6] G. Ananthakrishnan, A. K. Jayaraman, T. Trueman, S. Mitra, A. A. K, and A. Murugappan, *Suicidal Intention Detection in Tweets Using BERT-Based Transformers*. 2022, doi: 10.1109/ICCCIS56430.2022.10037677.

[7] I. Ameer, M. Arif, G. Sidorov, H. Gòmez-Adorno, and A. Gelbukh, "Mental Illness Classification on Social Media Texts using Deep Learning and Transfer Learning," *arXiv preprint*, 2022, doi: 10.48550/arXiv.2207.01012.

[8] U. Ahmed, J. C. W. Lin, and G. Srivastava, "Hyper-graph-based attention curriculum learning using a lexical algorithm for mental health," *Pattern Recognition Letters*, vol. 157, pp. 135–143, 2022, doi: 10.1016/j.patrec.2022.03.018.

[9] A. Kumar, K. Sharma, and A. Sharma, "Hierarchical deep neural network for mental stress state detection using IoT based biomarkers," *Pattern Recognition Letters*, vol. 145, pp. 81–87, 2021, doi: 10.1016/j.patrec.2021.01.030.

[10] R. Skaik and D. Inkpen, "Using Social Media for Mental Health Surveillance: A Review," *ACM Computing Surveys*, vol. 53, no. 6, pp. 1-31, Dec. 2020, doi: 10.1145/3422824.

[11] A. Le Glaz *et al.*, "Machine Learning and Natural Language Processing in Mental Health: Systematic Review," *Journal of medical Internet research*, vol. 23, no. 5, May 2021, doi: 10.2196/15708.

[12] T. Zhang, A. M. Schoene, S. Ji, and S. Ananiadou, "Natural language processing applied to mental illness detection: a narrative review," *NPJ Digital Medicine*, vol. 5, no. 1, pp. 1–13, 2022, doi: 10.1038/s41746-022-00589-7.

[13] R. Salas-Zárate, G. Alor-Hernández, M. A. Paredes-Valverde, M. del P. Salas-Zárate, M. Bustos-López, and J. L. Sánchez-Cervantes, "Mental-Health: An NLP-Based System for Detecting Depression Levels through User Comments on Twitter (X)," *Mathematics*, vol. 12, no. 13, pp. 1–30, 2024, doi: 10.3390/math12131926.

[14] A. Vaswani *et al.*, "Attention is all you need," *Advances in Neural Information Processing Systems*, pp. 5999–6009, 2017.

[15] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 2019, pp. 4171–4186.

[16] T. Wolf *et al.*, "Transformers : State-of-the-Art Natural Language Processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020, pp. 38–45, doi: 10.18653/v1/2020.emnlp-demos.6.

[17] C. M. Greco, A. Simeri, A. Tagarelli, and E. Zumpano, "Transformer-based language models for mental health issues: A survey," *Pattern Recognition Letters*, vol. 167, pp. 204–211, 2023, doi: 10.1016/j.patrec.2023.02.016.

[18] Q. un Nisa and R. Muhammad, "Towards transfer learning using BERT for early detection of self-harm of social media users," *CEUR Workshop Proc.*, vol. 2936, pp. 1059–1070, 2021.

[19] S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, and E. Cambria, "MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare," *2022 Language Resources and Evaluation Conference, LREC 2022*, pp. 7184–7190, 2022.

[20] H. Wang, X. Hu, and H. Zhang, "Sentiment analysis of commodity reviews based on ALBERT-LSTM," *Journal of Physics: Conference Series*, vol. 1651, no. 1, 2020, doi: 10.1088/1742-6596/1651/1/012022.

[21] S. Kaur, R. Bhardwaj, A. Jain, M. Garg, and C. Saxena, "Causal Categorization of Mental Health Posts using Transformers," *ACM International Conference Proceeding Series*, vol. 1, no. 1, pp. 43–46, 2022, doi: 10.1145/3574318.3574334.

[22] T. Zhang, A. M. Schoene, and S. Ananiadou, "Automatic identification of suicide notes with a transformer-based deep learning model," *Internet Interventions*, vol. 25, pp. 1–8, 2021, doi: 10.1016/j.invent.2021.100422.

[23] F. Haque, R. U. Nur, S. A. Jahan, Z. Mahmud, and F. M. Shah, "A Transformer Based Approach To Detect Suicidal Ideation Using Pre-Trained Language Models," in *2020 23rd International Conference on Computer and Information Technology (ICCIT)*, 2020, pp. 1–5, doi: 10.1109/ICCIT51783.2020.9392692.

[24] R. Z. Lumontod III, "Seeing the invisible: Extracting signs of depression and suicidal ideation from college students' writing using LIWC a computerized text analysis," *International Journal of Research Studies in Education*, vol. 9, no. 4, pp. 31–44, 2020, doi: 10.5861/ijrse.2020.5007.

[25] N. Masuda, I. Kurahashi, and H. Onari, "Correction: Suicide ideation of individuals in online social networks (PLoS ONE)," *PLoS ONE*, vol. 9, no. 1, Apr. 2014, doi: 10.1371/annotation/d589857d-b3c6-4a16-acfe-423f9bf529f1.

[26] J. Pestian, H. Nasrallah, P. Matykiewicz, A. Bennett, and A. Leenaars, "Suicide Note Classification Using Natural Language Processing: A Content Analysis," *Biomedical Informatics Insights*, vol. 3, p. BII.S4706, Jan. 2010, doi: 10.4137/BII.S4706.

[27] R. Haque, N. Islam, M. Islam, and M. M. Ahsan, "A Comparative Analysis on Suicidal Ideation Detection Using NLP, Machine, and Deep Learning," *Technologies*, vol. 10, no. 3, 2022, doi: 10.3390/technologies10030057.

[28] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, "Detection of Depression-Related Posts in Reddit Social Media Forum," *IEEE Access*, vol. 7, pp. 44883–44893, 2019, doi: 10.1109/ACCESS.2019.2909180.

[29] S. Pachouly, G. Raut, K. Bute, R. Tambe, and S. Bhavsar, "Depression Detection on Social Media Network (Twitter) using Sentiment Analysis," *International Research Journal of Engineering and Technology*, vol. 8, no. 1, pp. 1834–1839, 2021

[30] M. Stankevich, A. Latyshev, E. Kuminskaya, I. Smirnov, and O. Grigoriev, "Depression detection from social media texts," in *Selected Papers of the XXI International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2019)*, 2019, pp. 279–289.

[31] M. Manisha, A. Kodali, and V. Srilakshmi, "Machine classification for suicide ideation detection on twitter," *International Journal of Innovative Technology and Exploring Engineering*, vol. 8, no. 12, pp. 4154–4160, 2019, doi: 10.35940/ijitee.L3655.1081219.

[32] R. Sawhney, P. Manchanda, P. Mathur, R. Shah, and R. Singh, "Exploring and Learning Suicidal Ideation Connotations on Social Media with Deep Learning," in *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Oct. 2018, pp. 167–175, doi: 10.18653/v1/W18-6223.

[33] S. Ji, C. P. Yu, S. F. Fung, S. Pan, and G. Long, "Supervised learning for suicidal ideation detection in online user content," *Complexity*, no. 1, pp. 1–10, Jan. 2018, doi: 10.1155/2018/6157249.

[34] F. A. Acheampong, H. Nunoo-Mensah, and W. Chen, "Transformer models for text-based emotion detection: a review of BERT-based approaches," *Artificial Intelligence Review*, vol. 54, no. 8, pp. 5789–5829, 2021, doi: 10.1007/s10462-021-09958-2.

[35] V. Venek, S. Scherer, L. P. Morency, A. S. Rizzo, and J. Pestian, "Adolescent suicidal risk assessment in clinician-patient interaction: A study of verbal and acoustic behaviors," in *2014 IEEE Workshop on Spoken Language Technology, SLT 2014 - Proceedings*, Apr. 2014, pp. 277–282, doi: 10.1109/SLT.2014.7078587.

[36] E. Lin, J. Sun, H. Chen, and M. H. Mahoor, "Data Quality Matters: Suicide Intention Detection on Social Media Posts Using a RoBERTa-CNN Model," *arXiv preprint*, pp. 13–17, 2024, doi: 10.1109/EMBC53108.2024.10782647.

[37] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*. 2009.

[38] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. USA: Cambridge University Press, 2008.

[39] J. H. Martin and D. Jurafsky, *Speech and Language Processing*. Stanford, 2025.

[40] R. Feldman and J. Sanger, *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press, 2006. doi: 10.1017/CBO9780511546914.

[41] C. C. Aggarwal and C. X. Zhai, *Mining text data*, vol. 9781461432234. 2013, doi: 10.1007/978-1-4614-3223-4.

[42] F. Azuaje, I. Witten, and F. E, "Data Mining: Practical Machine Learning Tools and Techniques," *BioMedical Engineering OnLine*, vol. 5, pp. 1–2, Jan. 2006.

[43] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *arXiv preprint* no. 1, 2019, doi: 10.48550/arXiv.1907.11692.

[44] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations," *arXiv preprint*, pp. 1–17, 2019, doi: 10.48550/arXiv.1909.11942.

[45] P. He, X. Liu, J. Gao, and W. Chen, "DeBERTa: Decoding-enhanced BERT with Disentangled Attention," *arXiv preprint*, 2020, doi: 10.48550/arXiv.2006.03654.

BIOGRAPHIES OF AUTHORS

Merinda Lestandy is a researcher at the Department of Electrical Engineering, Faculty of Engineering, Universitas Muhammadiyah Malang (UMM), Malang, Indonesia. She received her Bachelor's and Master's Degrees from Electrical Engineering and Informatics Department, Universitas Muhammadiyah Malang, and Brawijaya University, Indonesia, in 2015 and 2018 respectively. Currently she is a senior lecturer for research fields of data mining, sentiment analysis, text mining, and its applications. She can be contacted at email: merindalestandy@umm.ac.id.

Abdurrahim is a graduate student of Electrical Engineering at the Universitas Muhammadiyah Malang (UMM) with the study program taken is Electrical Engineering, he focuses on telematics with the scope of sentiment analysis. Currently he is pursuing a Master's degree at the Islamic University of Indonesia, majoring in Informatics, focusing on research in the field of data science. He can be contacted at email: 22917002@students.uii.ac.id.

Amrul Faruq is an Electrical Engineer and Computer Science Engineer. He obtained Bachelor's and Master's degree in Electrical Engineering in 2009 and 2013, from Universitas Muhammadiyah Malang and Universiti Teknologi Malaysia, respectively. His Ph.D. obtained from the Malaysia-Japan International Institute of Technology (MJIIT), Universiti Teknologi Malaysia, Kuala Lumpur. His research interests about computational data science and optimization algorithms. He can be contacted at email: faruq@umm.ac.id.

Muhammad Irfan was born in Mojokerto, Indonesia in 1966. He graduated in 1991 with a Bachelor of Engineering degree, from the Department of Electrical Engineering, Brawijaya University Malang, and a Master of Engineering in 2000 from the Department of Informatics, Sepuluh Nopember Institute of Technology (ITS), Surabaya. Currently, he is a senior lecturer at the University of Muhammadiyah Malang (UMM) and is active in Research and Community Service. He is currently pursuing his doctoral research program at the Razak Faculty of Technology and Informatics, Universiti Teknologi Malaysia. His research interests are in renewable energy and Computer Engineering. He can be contacted at email: irfan@umm.ac.id.