Peer-Review

Chakka, Lakshmi. 2025. "Data Requirements for Suicidal and Depressive Thought Detection from RedditTM with Recurrent Neural Networks." *Journal of High School Science* 9 (3): 323–44. https://doi.org/10.64336/001c.143549.

The only reason I do not give this a reject rating is that I may have misunderstood the author's premise, hence I will post my comments below with a chance for the author to respond.

1.If the dataset was the 'suicidewatch' subredditt; is it not a given that the persons who post there have suicide thoughts or ideation? What then, is the point of your ML to recognize if a post on that subredditt has words associated with suicide thoughts? Will it obviously not have words associated with these thoughts? Please explain and describe in the manuscript.

2.Where is the ground truth? If all you examined was this subredditt, how do you know how many of those posts were NOT associated with suicide thoughts? Why would a person post on that particular designated subreddit if they did not have suicide thoughts? Please discuss in the manuscript.

3.It is obvious that the larger the dataset, the better the prediction accuracy. Why do you need an ML to prove the obvious?

4.There are no reported F1 values, precision or recall values reported. There is no confusion matrix reported; niether is there an AUC or a PRAUC reported. No hyperparameters are reported and no mention is made of how hyperparameters wer optimized. Are any of these words reported collinear? i.e. what is the probability that the word X is correlated (frequently occurs) with the word Y and both appear as features in the prediction analysis? Was a multicollinearity analysis with Pearson coefficients performed? If yes, please report VIF values, if not please justify. Also present a correlogram or a correlation matrix. Was any X-fold cross-validation performed? Why or why not? How was the class imbalance between the frequency of the feature words processed? SMOTE? data augmentation?

5.I need to see this model tried out on Reddit posts that are not ' suicidewatch' subredditts. I will need to see accuracy etc….. see point 4.

I cannot review this manuscript unless - as communicated previously to the author - they submit a separate word doc with the questions/comments of the reviewer and their answer to each question/comment. The responses must also include how and where that response is addressed in the manuscript. When I receive the manuscript with the responses, I will resume review.

Thank you to the reviewer for the detailed feedback and comments. I sincerely apologize for the delay in responding. The insights provided were very helpful, and the manuscript has been updated to address them.

REVIEWER COMMENTS:

1. **If the dataset was the 'suicidewatch' subredditt; is it not a given that the persons who post there have suicide thoughts or ideation? What then, is the point of your ML to recognize if a post on**
   that subredditt has words associated with suicide thoughts? Will it obviously not have words associated with these thoughts? Please explain and describe in the manuscript.
   Regarding the dataset composition, I would like to clarify that the "non-suicidal" category consists of posts taken from the "depression" subreddit, in addition to posts from the "SuicideWatch" subreddit. Addressing your first comment, while the SuicideWatch subreddit was created for discussions around suicide-related topics, not all posts indicate suicidal thoughts or ideation. Some users may post to discuss

suicide-related topics in general and how might society attempt to prevent these issues but not necessarily be suicidal themselves. Additionally, some posts may reference suicide in a non-serious or joking way, or they may be mistakenly posted in that specific subreddit. These examples highlight that not all content in the "SuicideWatch" or "depression" subreddits includes language clearly associated suicide or non-suicidal content. I would also like to note that one of the primary goals of this study is not limited to suicidal detection on Reddit. It also aims to develop a machine learning model that can differentiate between suicidal and non-suicidal, depressive content in many different contexts. While Reddit provides a valuable dataset for training, the broader goal is to create tools that can generalize to many applications which I have discussed in my conclusion section. I highlighted changes related to this comment in green below.

2. Where is the ground truth? If all you examined was this subredditt, how do you know how many of those posts were NOT associated with suicide thoughts? Why would a person post on that particular designated subreddit if they did not have suicide thoughts? Please discuss in the manuscript. The ground truth for this study comes from the pre-labeled dataset provided by the original creators. Each post in the dataset was labeled as either "suicidal" or "non-suicidal" based on patterns in text and contextual information present in the text. These labels were assigned during the dataset creation process, ensuring that the distinction between the two categories was established before the models were trained and tested. Since this comment is related to concerns in the first comment, I also highlighted changes related to this in green below.
3. It is obvious that the larger the dataset, the better the prediction accuracy. Why do you need an ML to prove the obvious? In my paper, I have mentioned that larger data sample sizes typically tend to have better results than smaller sample sizes. The primary goal of my study, however, was not just to confirm this, but also identify the ideal sample size that balances accuracy and resource efficiency. I wanted to find out how much data is really necessary for the model to perform well. I believe this would really benefit researchers in the future because collecting and managing large datasets can be costly and time-consuming and by pinpointing the optimal amount of data required for optimal performance, this research provides a practical solution. Changes I have made related to this comment are highlighted inorange below to further clarify and emphasize this.
4. There are no reported F1 values, precision or recall values reported. There is no confusion matrix reported; niether is there an AUC or a PRAUC reported. No hyperparameters are reported and no mention is made of how hyperparameters wer optimized. Are any of these words reported collinear? i.e. what is the probability that the word X is correlated (frequently occurs) with the word Y and both appear as features in the prediction analysis? Was a multicollinearity analysis with Pearson coefficients performed? If yes, please report VIF values, if not please justify. Also present a correlogram or a correlation matrix. Was any X-fold cross-validation performed? Why or why not? How was the class imbalance between the frequency of the feature words processed? SMOTE? data augmentation? The manuscript has been updated to include precision, recall, F1 scores, a confusion matrix, and PRAUC/AUC values to evaluate the model's performance. Hyperparameter details, including the learning rate, batch size, and training epochs, have also been added. A formal multicollinearity analysis was not conducted because the focus of this research is not on feature relationships but rather on evaluating dataset size and model architecture for performance optimization. Stratified K-

fold cross-validation was employed during training. Changes I have made related to this comment are highlighted in blue below

5. I need to see this model tried out on Redditt posts that are not ’ suicidewatch’ subredditts. I will need to see accuracy etc….. see point 4. This was a mistake on my end - while the dataset was named differently, the data actually consisted of posts from both the “SuicideWatch” and “depression” subreddits of Reddit and aimed to distinguish between those two. Changes I have made related to this comment are highlighted in green below.

____

Thank you for attempting to address my comments; some of which have not been addressed. As a result, you have not made a quantitative convincing case for your algorithm. Please see the comments below for deficiencies that have not yet been addressed.

1.The data on F1, accuracy, recall and precision with the confusion matrices that needs to be presented is for models that used signifiantly less data. You have only presented these values for the model that used the entire dataset (232074). Please present all these values for the model that used 1%, 2% of the dataset. If you used 2% of the dataset you would have approximately 4642 data points. I am assuming you would still have an 80:20 split ratio; i.e. 3714 training points and 928 test points. Where are these metrics (F1, accuracy, recall, precision) for these data points in the manuscript? In addition, if your claim about using only 2% of the dataset (without significant decrease in accuracy) is true, then you should have been able to use 3714 training points with the rest of the total dataset 228360 points as the testing set and still obtain the same accuracy. Where is this data presented in the manuscript.

2.For the K-fold validation, I would have thought you would have used a 50-fold validation; so as to provide evidence that no matter which 2% of the datapoints you used in the entire dataset, you ended up with similar accuracy, precision, recall, F1 score…… Please present this data with the average of these values across the 50-sets of data and the standard deviation.

3.Your model does not seem to use contextual algorithms. Hence, I would like to see more testing datasets (at least 3, the more the better) made up of ‘normal’ social media messages (that are neither depressed or suicidal). This is because, your model may flag the following sentences as being suicidal : for example:

a. “Im like, don’t you know you can’t get this mad at your friends?”

b. “Iam like, I don’t even want to know dude!”

c. “Iam like, get a life dude!”

d. “Ive lived long enough to know that you can’t get everything you want”.

etc. etc. Note that since your algorithm only ranks certain words in importance, all the sentences above are likely to be flagged as being suicidal, even though they are not. Hence, please run your algorithm of different datasets from Reddit that have nothing to do with either depression or suicide. Please make sure you include messages from teens (in specific redditts) since you will likely encounter these kind of messages in teen social media. Please present the usual metrics (F1, accuracy, precision, recall.etc.) for these datasets as well.

____

Thank you to the reviewer for the detailed feedback and comments. The insights provided were very helpful, and the manuscript has been updated to address them.

REVIEWER COMMENTS:

6. The data on F1, accuracy, recall and precision with the confusion matrices that needs to be presented is for models that used signifiantly less data. You have only presented these values for the model that used the entire dataset (232074). Please present all these values for the model that used 1%, 2% of the dataset. If you used 2% of the dataset you would have approximately 4642 data points. I am assuming you would still have an 80:20 split

ratio; i.e. 3714 training points and 928 test points. Where are these metrics (F1, accuracy, recall, precision) for these data points in the manuscript? In addition, if your claim about using only 2% of the dataset (without significant decrease in accuracy) is true, then you should have been able to use 3714 training points with the rest of the total dataset 228360 points as the testing set and still obtain the same accuracy. Where is this data presented in the manuscript? In my updated manuscript, I have presented the values (confusion matrix, etc) necessary to understand the model's performance using smaller percentages of the dataset. I have additionally tested my program again using a small amount of data for training (2% for example) and then using the rest as testing data. I highlighted changes related to this comment in orange below.

7. For the K-fold validation, I would have thought you would have used a 50-fold validation; so as to provide evidence that no matter which 2% of the datapoints you used in the entire dataset, you ended up with similar accuracy, precision, recall, F1 score…… Please present this data with the average of these values across the 50-sets of data and the standard deviation for all the metrics. In my updated manuscript, I have presented the average of the values across all parts of the data with the standard deviation. I highlighted changes related to this comment in orange below.
8. Your model does not seem to use contextual algorithms. Hence, I would like to see more testing datasets (at least 3, the more the better) made up of 'normal' social media messages (that are neither depressed or suicidal). This is because, your model may flag the following sentences as being suicidal : for example:
   - "Im like, don't you know you can't get this mad at your friends?"
   - "Iam like, I don't even want to know dude!"
   - "Iam like, get a life dude!"
   - "Ive lived long enough to know that you can't get everything you want".

   etc. etc. Note that since your algorithm only ranks certain words in isolation in importance, all the sentences above are likely to be flagged as being suicidal, even though they are not. Hence, please run your algorithm of different datasets from Redditt that have nothing to do with either depression or suicide. Please make sure you include messages from teens (in specific redditts) since you will likely encounter these kind of messages in teen social media. Please present the usual metrics (F1, accuracy, precision, recall.etc.) for these datasets as well. The manuscript has been updated after including additional posts from general subreddits not related to suicide or depression.Since I was unable to find a general dataset with more than 60,000 posts, I adjusted the number of suicidal and depressive posts fed into the model to match that amount. The total dataset size is now 180,000, with each class equally represented. I've also updated my results after these changes. I highlighted changes related to this comment in green below.

Thank you for addressing my comments. However, I am still skeptical since you have not addressed point 3 (see below).
Hence, I would like you to run your algorithm on the attached set of 30 sentences (none of which can be construed to be 'suicidal posts' Please present your results in the manuscript. I acknowledge that the sample size is too small;

however, if your algorithm is contextual, it should be accurate enough even with this small a sample size.
See attached file for sentences.

Your model does not seem to use contextual algorithms. Hence, I would like to see more testing datasets (at least 3, the more the better) made up of 'normal' social media messages (that are neither depressed or suicidal). This is because, your model may flag the following sentences as being suicidal : for example:

a. "Im like, don't you know you can't get this mad at your friends?"
b. "Iam like, I don't even want to know dude!"
c. "Iam like, get a life dude!"
d. "Ive lived long enough to know that you can't get everything you want".

etc. etc. Note that since your algorithm only ranks certain words in importance, all the sentences above are likely to be flagged as being suicidal, even though they are not. Hence, please run your algorithm of different datasets from Redditt that have nothing to do with either depression or suicide. Please make sure you include messages from teens (in specific redditts) since you will likely encounter these kind of messages in teen social media. Please present the usual metrics (F1, accuracy, precision, recall.etc.) for these datasets as well.

Thank you to the reviewer for the detailed feedback and comments. The insights provided were very helpful, and the manuscript has been updated to address them.

REVIEWER COMMENTS:

9. Thank you for addressing my comments. However, I am still skeptical since you have not addressed point 3 (see below). Hence, I would like you to run your algorithm on the attached set of 30 sentences (none of which can be construed to be 'suicidal posts' Please present your results in the manuscript. I acknowledge that the sample size is too small; however, if your algorithm is contextual, it should be accurate enough even with this small a sample size.

See attached file for sentences. In my updated manuscript, I have done my best to address this comment. I ran my model on the 30 normal sentences provided and included the results in my manuscript. My model, unfortunately, was not able to correctly classify most of them, with only 11 (sentences 2, 6, 9, 11, 12, 13, 14, 15, 22, 26, 29 in the given file) out of the 30 identified as normal. Because of these results, I considered using datasets using more contextually complex language so that this could be used for training data, however, I could not find any publicly available dataset that would be able to capture what I was looking for. Due to these data access and labeling reliability limitations, I addressed this issue as part of the Discussion section. I hope this current version addresses your concerns as much as possible given the scope of my study.

Thank you again for your feedback. Changes have been highlighted in yellow below.

Your model does not seem to use contextual algorithms. Hence, I would like to see more testing datasets (at least 3, the more the better) made up of 'normal' social media messages (that are neither depressed or suicidal). This is because, your model may flag the following sentences as being suicidal : for example:

- "Im like, don't you know you can't get this mad at your friends?"
- "Iam like, I don't even want to know dude!"
- "Iam like, get a life dude!"
- "Ive lived long enough to know that you can't get everything you want".

etc. etc. Note that since your algorithm only ranks certain words in importance, all the sentences above are likely to be flagged as being suicidal, even though they are not. Hence, please run your algorithm of different datasets from Redditt that have nothing to do with either depression or suicide. Please make sure you include messages from teens (in specific redditts) since you will likely encounter these kind of messages in teen social media. Please present the usual metrics (F1, accuracy, precision, recall.etc.) for these datasets as well.

Thank you for attempting to address my comments. I think we can make this work if you put these post-hoc comments down as an appendix in the manuscript.

You will need to submit a word docx of your manuscript with formatting consistent with the Journal's guidelines: one column text, 12 Times New Roman font. References numbered sequentially in the manuscript in curved brackets. Reference section in APA format. If > 6 authors, the first 6 must be listed followed by an et al., if less than 6 authors, all authors must be listed. Each reference must have a live link.

Please include the word docx in the next iteration so that I can forward for copyediting and potential approval.

Thank you for attempting to address my comments. I think we can make this work if you put these post-hoc comments down as an appendix in the manuscript. You will need to submit a word docx of your manuscript with formatting consistent with the Journal's guidelines: one column text, 12 Times New Roman font. References numbered sequentially in the manuscript in curved brackets. Reference section in APA format. If > 6 authors, the first 6 must be listed followed by an et al., if less than 6 authors, all authors must be listed. Each reference must have a live link.

Please include the word docx in the next iteration so that I can forward for copyediting and potential approval.

Thank you to the reviewer for the detailed feedback and comments for all of the manuscripts that I have submitted. I learned a lot from the process and am grateful for this experience. My manuscript has been updated to address the final comments and formatting requirements. I'd just like to note that some of the references that I have did not have a DOI link so I provided the original link in the references.

Accepted