AI-Driven Sentiment Analysis of Social Media for Mental Health Indicators

Mason Conkel and Sudman Sakib

Introduction: Increase use of social media has been attributed to a decline of mental health across users (1 and 2). To ameliorate this, researchers attempt to leverage the sentiment of a message to predict the mental status of the individual who posted. As social media is considered big data, this tasks is assumed within the domain of natural language processing (NLP). NLP models can leverage the vast quantities of social media posts with labels given by experts to predict the mental status of the posts made by individuals. With this, actions can be taken to benefit those that post the data, such as new recommendations for what content to consume. This work aims to leverage dual NLP models and construct a metric for model recommendations based on bot model outputs.

Related Work: Literature related to this topic can be discussed in two trust. Firstly, literature varies based on data used in training. (3) leverages social media containing emotion-carrying hashtags to create a dataset on which they train models. In contrast, (4) acquires data from sentences in suicide notes. For simplicity, we utilize pre-labeled datasets composed of social media posts to fit the issue.

The model used is the most researched aspect of literature on sentiment analysis. (3) deploys contemporary deep neural networks while (6-9) experiment with NLP models such as recurrent neural networks (6, 8 and 9), long short-term memory (LSTM) (6, 7 and 8) and transformer-based models (BERT) (6 and 8). These provide a foundation on which models can be built while providing examples for benchmarks: accuracy, precision, recall and f1-score. These can increase reliability and certainty of model performance and provide literature against which we can prepare. However, none of these seek to leverage multi-class labels for deeper analysis of a message to reinforce prediction certainty.

Datasets and Labels: Three datasets are used in the creation and testing of the system of models. [1] supplies the foundation for which a model may be trained to predict the emotion sentiment from a text. This means each text is given a label from a set of emotions: happy, sad, angry, etc. Upon this was reduced from the eight labels in the original to the 6 labels of anger, fear, joy, neutral, sadness, and surprise. The process by which disgust and shame were dropped will be explained in later sections.

[2] provides the first application of multi-model attention. This dataset offers mental health labels to text such as depressed, suicidal, normal, etc. These labels could be used to train a second model that will provide a mental health status to an individual. However, as models are prone to a degree of error, the first model can used to make predictions on the second dataset and a cross-tabular matrix can be created that underscores the intersection of mental health and emotional sentiment. This table will be discussed further in our results.

[3] is a dataset that neither model was trained on. This dataset is composed of text entries with a binary classification scheme for a label that designates the text as normal or depressed. Using this as a validation set, both models can provide predictions on the text and the performance of both can be analyzed with expert judgment to determine how the models perform.

Metrics: The most pertinent metrics to model performance are chosen as accuracy, f1 score and confidence. Accuracy is a common metric on which to evaluate models as it clearly describes the relationship between between true labels and predicted labels. This is helpful for individual model training as we desire models to perform as close to the experts that labeled the datasets as is possible when seeing new data. It is also the simplest metric to explain as the percent of correct predictions.

For this same reason, F1 score is a good metric to determine model performance as it compares true positives (predicted = label) to the combination of true positives, false positives (if predicted is label and wrong, then label receives a false positive) and false negatives (if prediction is wrong, the correct label receives a false negative). The corresponding equation follows.

$$F1 = \frac{2 \times TP}{2 \times TP + FP + FP} = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

The final metric we use for determining model effectiveness, is confidence. For LLMs, the output is usually determined after a softmax function by taking the resulting maximum argument. Confidence is obtained by taking the maximum value instead of the maximum argument. As the softmax operation changes the outputs to a interval of (0, 1) in which the sum of all outputs equals to 1, it can be determined that a high confidence corresponds to how sure the model is in its decision. This can be useful when deciding to perform an operation based on the model prediction. As that is what we do with these values, our operation and usage of confidence will be described in further detail later.

Models: We deploy and train three models to compare their efficiency. Two of these models utilize state-of-the-art architectures and use pre-trained fine-tuned weights that we adapt for sentiment analysis. These architectures are BERT and GPT2. We use the standard configurations and fine-tuned weights as provided by PyTorch and add a linear layer with the number of neurons equal to the number of labels. Similarly, tokenizers are retrieved from the BERT and GPT2 pre-trained models that correspond to their models. These are used in tandem with the declared models and trained on our datasets. Tests showed that a text length of 256 with 10 epochs were best for both models as larger representations did not add value to the end result.

The final model is a standard transformer encoder similar to the BERT model but with significantly reduced size of 9,319,942 parameters. This model is created using an embedding layer with a vacab size of 30,000 and an embedding dimension of 256 accounting for 7.68 million parameters. Next, a positional encoding layer using a tensor of max text length (256) and the embedding dimension produces an additional 65,536 parameters. With a model dimension of 256, 6 heads and a feed forward layer of 512, the total transformer parameters are 1,572,864. The final classifier layer uses the embed dimensions and number of classes for the last addition of 1,542 parameters. The total, approximately 9.3 million parameters, are one tenth the size of the smallest BERT (~110 million) and GPT2 (~117 million). We argue that this make the model more lightweight and provides better potential for deployment.

Results: As discussed, the first model is used to provide predictions over the second dataset. Next, the labels from that dataset and the predictions are used to construct a cross-tabular table that can be used too derive key information that can reinforce model predictions. These processes are done for all models and GPT2 the first results that illustrate this process.

GPT2: The emotional sentiment model produced a testing accuracy of 65.54% and an F1 of 0.6535 for an overall fair performance rating on the test set. The following cross-table was constructed from the mental health dataset.

Mental State	Anxiety	Bipolar	Depression	Normal	Personality Disorder	Stress	Suicidal
Predicted Emotion							
Anger	400	502	2258	1824	257	610	1592
Fear	1500	380	1822	2501	206	477	1408
Joy	1329	1252	4463	5007	351	1027	2436
Neutral	22	6	102	844	5	13	77
Sadness	551	594	6382	3762	244	429	4888

Surprise	39	43	377	2405	14	31	179

From this table, relationships between emotion and mental status can be created based the emotion model. For example, two major mental health issues often associated with social media are depression and suicide. Using this chart, a relation can be made between each of these and their primary contributors. For example, it is common knowledge that melancholy is associated with depression. From this chart, we can take the intersection of depression and sadness as 6382 predicted occurrences. Then, we can compare this to the total predictions made on the depression label, which is 15404 occurrences. Through probability theory, we can state that, for GPT2, the probability of depression given sadness P(Depression| Sadness) is approximately 41.43%. Additionally, the probability of suicidal tendencies given sadness P(Suicidal| Sadness) is approximately 46.20%. Through this information, we can also remove emotion predictions that are not chosen as often. For example, shame and disgust had low contributions to each label and would not contribute to the final decision as well as the selected six.

The next phase is to train the second model on the mental status data, on which GPT2 scored an accuracy of 76.98% and an F1 of 74.19% showing above average training results. With both models ready, predictions were made on the third, unseen dataset and samples are provided:

"nothing look forward lifei dont many reasons keep going feel like nothing keeps going next day makes want hang myself"

Label: Health Issues – Emotion: Sadness; 99.98% - State: Suicidal; 68.22%

From this text, an expert would produce a label that corresponds with the predictions from both GPT2 models. However, automating this task requires more certainty in the operational capabilities. From the predicted labels, we ascertain that the sadness prediction has a confidence of 99.98% and the suicidal prediction has a confidence of 68.22%. With the aforementioned probability of suicidal tendencies given sadness being 46.20%, we can create a metric that actions can be based on. In this paper, we propose the following equation for sentiment awareness:

$$SAware = 1/\sqrt{2} \times P(State \mid Emotion) * RSS(Conf_{Emotion}, Conf_{State})$$

Given this, the SAware would be determine using the root sum square of both confidences over the square root of two multiplied by the probability of the stat given the emotion. This would result in a value of 39.54%. Then, we can assign a threshold value over which we conduct an action such that SAware > $SAware_{Thres}$ = Execute. In the case of this paper, a threshold of 25% (all values at 50%) could suffice, meaning the previous task would correctly flag the post and execute a response. Using this strategy, we can observe a second instance.

"finally upgarded grandma bought new phone beneath using s yearsd"

Label: Normal – Emotion: Neutral; 96.06% - State: Normal; 81.21%

With the P(Normal | Neutral) being 5.16%, the resulting sentiment awareness is 4.59%. In this circumstance, the model fails to execute an operation, but this is acceptable error as normal behavior is not what the model was intended to counter. Therefore, this can be an example of a successful fail. In fact, according to the equation and the overall predictions of the neutral category, any prediction made

by the mental status model is overwritten when the emotion model predicts neutral status. Given that such examples may be edge cases, this can be considered a preferred outcome of model training.

"rabbit died rabbit named thumper died fell garden pot broke neck really nice bunny fearless scared anyone add insult injury happened right birthday idk posting sad"

Label: Normal – Emotion: Sadness; 99.98% - State: Suicidal; 78.45%

For this final example, we highlight a case in which the original dataset label differs from model predictions. With the P(Suicidal | Sadness) ane the current values, the SAware becomes 41.52% and would result in direct action taken. As many keywords in this post have meanings that could result in the model confusion, the resulting error is one that could be evaluated in future work.

BERT: The first model was able to achieve a test accuracy of 74.34% and an F1 of 0.7393. It is expected that BERT should be more accurate than GPT2 on classification tasks without Chain-of-Though prompting to push GPT models. The following cross-table was generated for BERT.

Mental State	Anxiety	Bipolar	Depression	Normal	Personality Disorder	Stress	Suicidal
Predicted Emotion							
Anger	113	259	1589	1352	136	395	1274
Fear	2252	474	2473	999	204	548	1742
Joy	176	236	546	4120	51	234	272
Neutral	41	58	91	2735	33	23	69
Sadness	1246	1734	10610	5396	625	1375	7217
Surprise	13	16	95	1741	28	12	78

For this table, the P(Depression | Sadness) is now 68.87% meaning that the occurrences in which both sadness and depression are predicted are weighed heavier by BERT than by GPT2. The second model is trained and achieves an accuracy of 82.93% and an F1 score of 0.8289. Then, we deploy the models on the third set and analyze their performance.

The first post listed under GPT2 achieved the same labels for BERT, however the confidence scores were 88.76% and 74.04%, respectively. Using a P(Suicidal | Sadness) of 67.75% gives an SAware of 55.38%. We argue that for this text, the model is more aware of the intended sentiment and an action would be taken.

The second post about "grandma" was predicted as sadness by BERT with a confidence of 55.48%. However, the mental status predicted normal with a confidenc of 99.98%. With the P(Normal | Sadness) equal to 33.02%, the final SAware is 26.69%. If a $\text{SAware}_{\text{thresh}}$ is 25%, then this model exceeds it by 1.69% and we would assume that action could be taken. Since the emotion is predicted as sad and the mental state as normal, minor changes to predictive algorithm could be made to brighten the mood. For this instance, the GPT2 model performed better during emotional prediction, but both mental status models performed optimally.

The final post about the "rabbit" was predicted as sadness with a confidence of 99.29% and normal with a confidence of 99.96%. The SAware for this is 32.89% and a comparison can be made to the last text. Before, we establish a threshold of 25%. This would result in both posts receiving the same treatment for sadness with normal health condition. However, adjusting $\text{SAware}_{\text{thresh}}$ to 30% would fix this issue and only the third post would receive action while the second would be ignored for lack of confidence.

We can also examine the following example. For futher performance analysis.

"family friends one talk toand people internet offer talk them also severe trust issues paranoia cant develop healthy relationship anyone unless comfortable someone tell really feel basically im traped mind able tell anyone sick really want die"

Label: Health Issues – Emotion: Fear; 99.97% - State: Depressed; 99.96%

For this example, the P(Depressed | Fear) is 16.05%. Therefore, the SAware is 16.05%. This result comes from the low value for the probability overwritting the near perfect confidence scores. An expert would say both models were correct and the additional percentage requirement may hinder the performance in these cases.

"deserve liveif died right noone would carei real friendsi always start conversations get dry responses i feel comfortable around females emotional abuse mom put left usi never find love i keep getting reminded everyday failure disappointment parents compared siblingsi first suicidal thought afraid grades good enough parents this probably end soon"

Label: Health Issues – Emotion: Fear; 75.11% - State: Suicidal; 98.51%

P(Suicidal | Fear) is 16.35% and the SAware is 14.33%. Again, the model would not predict this correctly with an $SAware_{thresh}$ of 25%. However, knowing that the prediciton of fear gives generally lower probability values than sadness, we can adjust the threshold again. Earlier we stated an adjustment to 30% could help filter sadness error while this example could benefit by adjusting to 10%. Therefore, we can establish emotion-depending thresholding such that $SAware_{Sadness}$ = 30% and $SAware_{Fear}$ = 10%. Now, we can provide reliable predictions using probability and confidence.

Custom: The custom small matrix produces an accuracy of 58.94% and an F1 score of 59.00%. This places the model just below GPT2 in potential performance and its decisions alone may prove unreliable for decision making. This produces the following cross table.

Mental State	Anxiety	Bipolar	Depression	Normal	Personality Disorder	Stress	Suicidal
Predicted Emotion							
Anger	737	471	2518	4111	260	570	1751
Fear	1180	486	1249	2166	137	329	970
Joy	387	284	902	3149	62	408	445
Neutral	1345	1457	9379	1073	573	1202	6545
Sadness	171	78	1215	3370	42	75	903
Surprise	21	1	141	2474	3	3	38

From this data, we find our common indicator P(Depression | Sadness) is now 7.89%. This will mute the corresponding values, but our techniques discussed earlier could enhance the value of this model further. We will explore this compared to earlier examples.

For the first text concerning suicidal thoughts, the first model provides an emotion label of joy with a confidence of 59.65% while the second marks suicidal with a confidence of 44.14%. The P(Depression | Joy) is 5.8% and the resulting SAware value is 3.07%. The second text concerning a phone upgrade received predictions of neutral (97.16%) and normal (91.39%). With P(Normal |

neutral) of 6.5%, the corresponding SAware is 6.19%. Finally, for GPT2 comparison, the final message regarding a pet death received predictions of sadness (99.96%) and depression (37.89%). With the probability of 7.89%, the SAware is 5.96%.

These imply that the average sentiment awareness of the model is much lower than observed in the BERT and GPT2 models. For this reason, we can shift the $SAware_{thresh}$ to sub 10% values. However, we need to observe other text that the BERT used to determine specific thresholds.

The first of BERT texts concerning family friends provided labels of anger (99.25%) and stress (30.15%). With P(Stress | Anger) of 22.85%, the SAware would be 16.16%. The final text about other suicidal thoughts that provided label of anger (99.99%) and suicidal (86.67%). With the value of P(Suicidal | Sadness) at 9.48%, the final SAware is 7.93%.

Concluson: Using SAware, smaller and typically less reliable models could still be used for predictions despite their performance when compared with state-of-the-art models. However, when deployed with a state-of-the-art model, such as BERT, it can inform on the model's decision and provide a value on which actions can be automated with reliable outcomes. Their predictions can be used for automated responses after initial setup of the SAware threshold. Future research can explore the use of this metric for the development of a new network based on confidence or defining futher reliability of the metric.

Datasets:

[1] https://www.kaggle.com/datasets/rikinzala/emotion-dataset-raw
[2] https://www.kaggle.com/datasets/suchintikasarkar/sentiment-analysis-for-mental-health
[3] https://www.kaggle.com/datasets/reihanenamdari/mental-health-corpus

References:

1. Braghieri, Luca, Ro'ee Levy, and Alexey Makarin. 2022. "Social Media and Mental Health." *American Economic Review* 112 (11): 3660–93. DOI: 10.1257/aer.20211218
2. Chancellor, S., De Choudhury, M. Methods in predictive techniques for mental health status on social media: a critical review. npj Digit. Med. 3, 43 (2020). https://doi.org/10.1038/s41746-020-0233-7
3. Muhammad Abdul-Mageed and Lyle Ungar. 2017. "EmoNet: Fine-Grained Emotion Detection with Gated Recurrent Neural Networks." In Proceedings of the *55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 718–728, Vancouver,Canada. Association for Computational Linguistics.
4. Islam, M.S., Kabir, M.N., Ghani, N.A. et al. "Challenges and future in deep learning for sentiment analysis: a comprehensive review and a proposed novel hybrid approach". *Artif Intell Rev* 57, 62 (2024). https://doi.org/10.1007/s10462-023-10651-9
5. Wadawadagi, R., Pagi, V. "Sentiment analysis with deep neural networks: comparative study and performance assessment." *Artif Intell Rev* 53, 6155–6195 (2020). https://doi.org/10.1007/s10462-020-09845-2
6. Sayyida Tabinda Kokab, Sohail Asghar, Shehneela Naz. "Transformer-based deep learning models for the sentiment analysis of social media data." https://doi.org/10.1016/j.array.2022.100157
7. Kaur, G., Sharma, A. "A deep learning-based model using hybrid feature extraction approach for consumer sentiment analysis." *J Big Data* 10, 5 (2023). https://doi.org/10.1186/s40537-022-00680-6
8. Soumitra Ghosh, Asif Ekbal, Pushpak Bhattacharyya, "Chapter 2 - Natural language processing and sentiment analysis: perspectives from computational intelligence." https://doi.org/10.1016/B978-0-32-390535-0.00007-0

9. Iqra Ameer, Necva Bölücü, Muhammad Hammad Fahim Siddiqui, Burcu Can, Grigori Sidorov, Alexander Gelbukh, “Multi-label emotion classification in texts using transfer learning,” https://doi.org/10.1016/j.eswa.2022.118534