**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.

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 (\_\_\_final size\_\_\_). (\_\_\_model specifics\_\_\_).

**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.

The GPT2 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** |  |  |  |  | 2405 |  |  |
| 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:

Given this, the SA 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 SA > SAThres = 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 SA 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.

(\_\_\_custom model\_\_\_).

BERT displayed adequate performance for

**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>