Sentiment Analysis for Mental Health

Candidate No.: 021739
Department of Computer Science
University of Exeter, Exeter, UK

***Abstract*—This project aims to aid the support of mental health by applying text analysis and machine learning techniques to flag and classify mental health problems within social media posts and comments. This paper proposes a sentiment analysis model utilising principle component analysis transformed TF-IDF vectors within a support vector machine model, achieving a test accuracy of 75%.**

I. Introduction

According to the NHS, the number of people in contact with the *NHS funded secondary mental health, learning disabilities and autism services* has been increasing significantly every year, particularly since the advent of the Coronavirus pandemic in 2020 [1]. Furthermore, in 2023 the government proposed an additional £2.3B of funding for mental health services, highlighting it's importance [2].

Traditional applications of sentiment analysis typically revolve around binary classification problems, such as spam detection or determining the polarity of reviews. In contrast, identifying different mental health problems is a multi-class classification problem which is often challenging to implement. This project details the creation of a sentiment analysis model for use in identifying mental health problems in social media posts and comments.

Previous research into sentiment analysis for emotions and mental health has shown promising results using a range of techniques [3]. For instance, Chen et al (2018) [4] analysed posts from X (formerly Twitter) to identify and distinguish four different mental health issues. Their study utilised an 8-emotion lexicon algorithm called *EMOTIVE*, originally proposed by Sykora et al (2013) [5]. Lexicon algorithms are widely used in sentiment analysis tasks as it provides a way to extract lower-dimensional features from text which could then be put through a classifier model. A similar emotion lexicon algorithm will be explored in this project.

Recent advancements in sentiment analysis have increasingly incorporated the attention block mechanism introduced by Vaswani et al (2017) [6], which has seen success in mental health classification tasks [7], [8]. This type of approach enhances word vector embeddings by considering contextual relationships, leading to more accurate representation of text. A common implementation of this approach is BERT embeddings, which will also be explored in this project.

II. Dataset

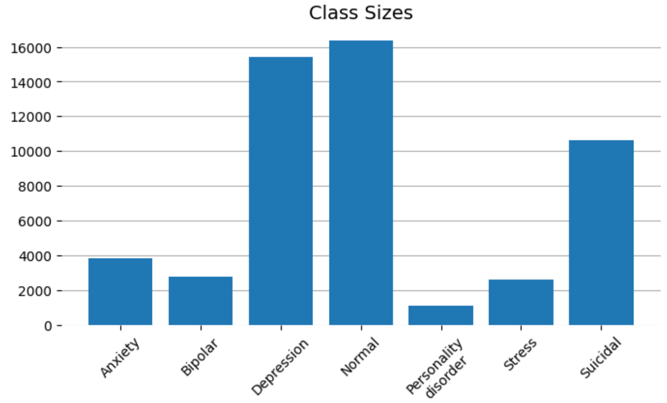

Fig. 1. A histogram showing the count for each class type in the *Sentiment Analysis for Mental Health* dataset.

The dataset used in this project was obtained via Kaggle.com, *Sentiment Analysis for Mental Health* [9], which compiles multiple sources into a 53,043 statement dataset containing six different mental health classes (and one control group). The dataset only contains two fields: the *statement* (text) and *status* (associated mental health condition). According to the datasets description [9], the statements consist of posts and comments from the social media platforms X and Reddit as well as chatbot training prompts. The mental health issues under consideration for this project are:

- Anxiety
- Bipolar
- Depression
- Personality disorder
- Stress
- Suicidal

Differentiating these conditions should be plausible although there may be overlap between anxiety and stress [10] as well as depression and suicidal [11].

Through some exploratory data analysis, it was revealed that the dataset suffers from one minor issue, class size imbalance - see figure 1. To mitigate this concern during model training and feature selection, an under-sampling technique will be imposed.

III. Feature Extraction

This project explored a range of different feature extraction techniques to find the best approach to the problem. Prior to feature extraction, some pre-processing was applied to the dataset including: removal of stop words, lowercasing, and lemmatisation.

Three different feature extraction techniques were tested with increasing complexity. Starting with the bag-of-words frequency vector via TF-IDF was intuitive as it only requires the provided dataset. NRC emotion lexicon was applied next to provide more meaningful insight into the data. Finally, as a demonstration of modern deep learning techniques, BERT embeddings were explored.

A. Bag-of-words TF-IDF Vectorisation

This first approach only requires the pre-processed statements and can be incredibly useful for gaining an initial insight into which words are strongly associated with each mental health issue. The TF-IDF vectorisation can be done through the Scikit-learn class `TfidfVectorizer`. However, applying this to our entire dataset unfortunately encounters a memory issue since, even after removing stop words and lemmatising, there are over 81,000 distinct words in our dataset making the array for this feature set of shape `(53043, 81077)` requiring 32GB of memory. To prevent this, and to mitigate the class imbalance issues, some undersampling was applied to only include 10% of the entire dataset. This undersampling was done such that the class sizes are now even.

Applying TF-IDF vectorisation to our subset yields an array of size `(5306, 21798)` (note that the second dimension here is dependent on which statements are sampled). By taking the class average of every dimension in this array, we can identify the top 5 key words associated with each class. Table I shows the results.

Issue	Key Words
Anxiety	anxiety, restless, feel, like, heart
Bipolar	bipolar, feel, like, know, get
Depression	feel, like, depression, want, life
Normal	go, want, morning, oh, really
Personality Dis.	like, people, avpd, feel, even
Stress	stress, like, feel, get, know
Suicidal	want, cannot, life, feel, like

TABLE I

TOP 5 KEY WORDS ASSOCIATED WITH EACH CLASS BASED ON 10% OF THE ENTIRE DATASET.

It is interesting to see some words which are common across all mental health classes but still give strong TF-IDF scores such as "*feel*" and "*like*"; these two in particular are likely common as a bigram used when people are describing how they feel, so perhaps this is expected and could help with distinguishing between mental health issues and the control group.

B. NRC Emotion Lexicon

Another feature extraction method explored was an emotion lexicon algorithm using the *NRC dataset* [12]. This dataset multi-classifies key words into one of ten different emotions: *anger, anticipation, disgust, fear, joy, negative, positive, sadness, surprise, trust.*

To convert each of our statements into a numerical vector representation, every pre-processed token was checked against the lexicon, and a sum was taken over all emotions associated with the words in a given statement. Each dimension (emotion) within the vectors can then be pseudo normalised by dividing them by the dataset mean count for that emotion, such that every statement is represented as a vector indicating relative presence of each emotion compared to the dataset.

Some preliminary analysis was done on this feature set to better understand what it represents. See appendix A for details on this; in summary, although it may be challenging to differentiate specific mental health issues from each other, they appear to distinguish well from the *normal* class.

C. BERT Embeddings

The final feature extraction technique used was BERT embeddings via a model provided by TensorFlow [13]. This model embeds each statement into a 512-dimensional space, utilising transformers to adjust each token vector based on the surrounding context. This gives a good representation of the true meaning behind each statement. It's important to note that the pre-processing previously mentioned was not applied to the text before being fed to the BERT model.

IV. RESULTS

This project explored three common approaches to sentiment analysis models: multi-layer perceptrons (MLP), support vector machines (SVM), and decision trees.

A. Model & Feature Selection

Once all of the different feature sets and models had been decided, a comparison of all nine combinations could be completed through 10-fold cross validation with stratified sampling. Given the restraints in computing the bag-of-words feature set and to maintain a reasonable runtime, this analysis was completed using the 10% subset described in section III.A. Furthermore, a final transformation was done to the bag-of-words feature set due to its significantly large dimensionality, PCA was used to reduce the dimension of this feature set to 1000, making it highly more manageable within models. The results of the 10-fold cross-validation are presented in table II.

Model	Features	Mean Acc.	Std. Acc.
MLP	Bag-of-words	0.564831	0.026346
SVM	**Bag-of-words**	**0.676777**	**0.015937**
Decision Tree	Bag-of-words	0.442888	0.023976
MLP	NRC emotion lexicon	0.355261	0.010639
SVM	NRC emotion lexicon	0.346591	0.017024
Decision Tree	NRC emotion lexicon	0.316248	0.013397
MLP	**BERT embeddings**	**0.592912**	**0.009698**
SVM	BERT embeddings	0.571047	0.017423
Decision Tree	BERT embeddings	0.374668	0.026446

TABLE II

RESULTS FROM THE 10-FOLD CROSS VALIDATION ON 10% OF THE DATASET. THE TOP TWO COMBINATIONS ARE IN BOLD.

B. Final Two Models

Interestingly, the most effective combination was the PCA transformed bag-of-words vector paired with an SVM achieving almost 68% accuracy with a standard deviation of less than

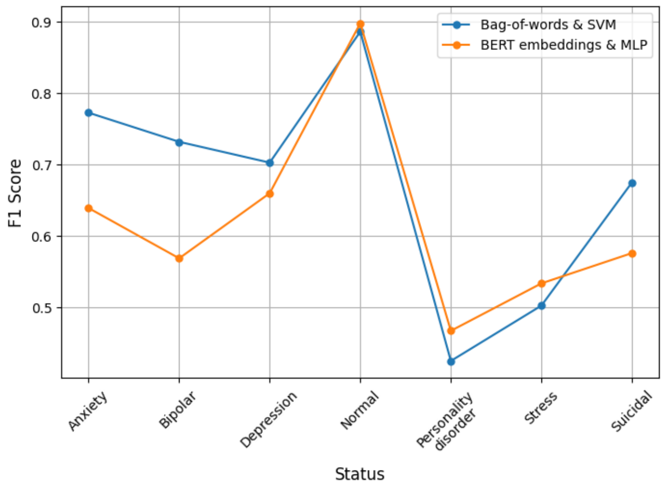

Fig. 2. F1 scores within each class for the two proposed models. The bag-of-words & SVM model performs better than the BERT embeddings & MLP model on all but three classes.

2%. The main difficulty with this feature set is its potential lack of generalisability, since the embeddings rely on dataset specific word distributions. For example, training the same model on a slightly different subset of our dataset might involve different key words, reducing repeatability. However, this issue could be mitigated through a pipeline which re-uses the trained `TfidfVectorizer()` and `PCA()` classes from this analysis rather than retraining them.

Consider the following pipeline:

1) Pre-process statements
2) Convert to TF-IDF vector using the vectoriser object pre-trained on 10% of the dataset - if words are not present in the dictionary for this object, they are ignored
3) Reduce dimensionality through the PCA object pre-trained on 10% of the dataset
4) Feed to SVM model

The next strongest performing model and feature set combination was an MLP with BERT embeddings. This is more what might have been expected given the complexity in BERT embeddings and the flexibility for capturing complex relationships in MLP models. Furthermore, the benefit to this combination is the simpler pipeline, no cleaning of statement is required since BERT embeddings work from raw text, and no dimensionality reduction is required.

Through some further cross validation and grid search, we can fine tune our two proposed models and do a fair comparison based on the entire dataset. See appendix B for details on the model fine-tuning via grid search. Using these models on our entire dataset, with 70% for training and 30% for testing, yields the following F1 scores for each mental health issue in figure 2.

Overall, the PCA transformed bag-of-words feature set paired with an SVM performed better than the BERT embeddings feature set paired with an MLP, achieving an overall test accuracy of 75% compared to 70%. Given that the goal of this project is to build a model to identify and flag mental health problems, a good metric to consider is the precision of the *Normal* class as this will indicate how often the model's classification as *normal* is correct. The SVM model achieved a normal precision of 83% compared to the MLP model's precision of 91%. This indicates that the SVM model might be missing some potential mental health issues which the MLP model is catching. Appendix C gives a full table of results for both models.

C. Results Discussion

The final strongest performing model created for this project was a Support Vector Machine (SVM) utilising PCA transformed TF-IDF vectors. The model effectively transforms each statement into a 1000-dimensional float vector based on patterns in word frequency. The use of PCA here is powerful as it allows for patterns in word counts to be considered and not just individual words.

The main challenge with this approach, as previously mentioned, is that it is highly dependent on which portion of the dataset is used for training the TF-IDF vectoriser and PCA model. Some further analysis could be done to verify the reproducibility of this approach. Perhaps given a more powerful machine, one could use the TF-IDF vectoriser on the entire dataset, this should indeed improve performance. An alternative approach could be to deterministically select which key words are considered in the TF-IDF vectorisation, this would be less dataset dependent and is more akin to the techniques used in lexicon algorithms.

One possible cause for the BERT embeddings approach not yielding as strong results could be due to the fact that BERT embeddings are a general way to encode the meaning of any statement. It's highly plausible that, since all of the statements in this dataset (besides the control class) are discussing a mental health problem, their BERT vectors were more similar than would have been preferable. Furthermore, there's a lot of redundant noise in these vectors which do not need to be considered for our task. An improvement to this approach, given more time and computing power, would be to fine tune the BERT embeddings so that they are more sensitive to different mental health issues.

V. Conclusion

This project has demonstrated how sentiment analysis techniques could be used to identify mental health problems in social media posts and comments. The final proposed model achieved a test accuracy of 75%.

It is quite refreshing to see a more classical machine learning approach out-performing the contemporary deep learning standards. Not only did the final feature extraction method make use of nothing more than the dataset itself - no external pre-trained models, but did so with an SVM rather than MLP. This contradicts some modern understanding of the current industry standards but highlights the importance of traditional approaches and demonstrates that sometimes the simpler approach is the better one.

References

[1] N. H. S. England, "Nhs," https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/mental-health-data-hub/dashboards, accessed: 2025-03-09.

[2] D. M. Team, "How we are supporting mental health services in england," https://healthmedia.blog.gov.uk/2023/06/09/how-we-are-supporting-mental-health-services-in-england/, accessed: 2025-03-09.

[3] A. Zunic, P. Corcoran, and I. Spasic, "Sentiment analysis in health and well-being: Systematic review," *JMIR Med Inform*, vol. 8, no. 1, p. e16023, Jan 2020. [Online]. Available: https://medinform.jmir.org/2020/1/e16023

[4] X. Chen, M. Sykora, T. Jackson, S. Elayan, and F. Munir, "Tweeting your mental health: An exploration of different classifiers and features with emotional signals in identifying mental health conditions," 2018.

[5] M. Sykora, T. Jackson, A. O'Brien, and S. Elayan, "Emotive ontology: Extracting fine-grained emotions from terse, informal messages," 2013.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[7] K. K. Patel, A. Pal, K. Saurav, and P. Jain, "Mental health detection using transformer bert," in *Handbook of Research on Lifestyle Sustainability and Management Solutions Using AI, Big Data Analytics, and Visualization*. IGI Global Scientific Publishing, 2022, pp. 91–108.

[8] M. F. Hadi, N. Sulistianingsih *et al.*, "Using sentiment analysis with bert and svm for detect mental health detection on social media," *SHIFANA: Journal of Digital Health Innovation and Medical Technology*, vol. 1, no. 2, pp. 54–63, 2025.

[9] S. Sarkar, "Sentiment analysis for mental health," https://www.kaggle.com/datasets/suchintikasarkar/sentiment-analysis-for-mental-health?resource=download, accessed: 2025-03-03.

[10] N. Daviu, M. R. Bruchas, B. Moghaddam, C. Sandi, and A. Beyeler, "Neurobiological links between stress and anxiety," *Neurobiology of Stress*, vol. 11, p. 100191, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352289519300438

[11] J. G. Keilp, M. F. Grunebaum, M. Gorlyn, S. LeBlanc, A. K. Burke, H. Galfalvy, M. A. Oquendo, and J. J. Mann, "Suicidal ideation and the subjective aspects of depression," *Journal of Affective Disorders*, vol. 140, no. 1, pp. 75–81, 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165032712000924

[12] S. Mohammad, "Nrc emotion lexicon," https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.html, accessed: 2025-03-07.

[13] TensorFlow, "Tensorflow — bert — kaggle," https://www.kaggle.com/models/tensorflow/bert/tensorFlow2/en-uncased-preprocess/3?tfhub-redirect=true, accessed: 2025-03-07.

Appendix

A

To understand the NRC emotion lexicon feature set more, figure 3 shows the average vector values within each emotion for every mental health issue.

From figure 3, it seems that the vectors for most mental health issues may be challenging to distinguish from each other although easy to distinguish from the *normal* class.

B

To fine-tune the bag-of-words with SVM model, the following parameter combinations of the `sklearn.svm.SVC` class were tested through 5-fold cross validation: `kernel:[linear, poly, rbf, sigmoid], C: [0.01, 0.1, 1, 10, 100]`. The results indicated that `kernel=rbf, C=1.0` was the strongest combination.

To fine-tune the BERT embeddings with MLP model, the following parameter combinations of the `sklearn.neural_networks.MLPClassifier` class were tested through 5-fold cross validation: `activation:[logistic, tanh, relu], hidden_layer_sizes: [(50,), (100,), (50, 2), (100, 2)]`. The results indicated that `activation=logistic, hidden_layer_sizes=(100,)` was the strongest combination.

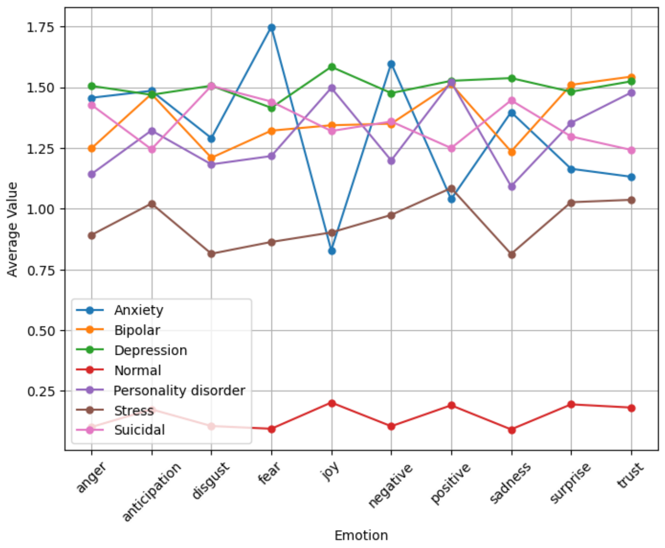

Fig. 3. Average vector values for every emotion, computed for the entire dataset.

C

Results from the analysis of the two final models on the entire dataset are presented in tables III and IV for the SVM and MLP models respectively.

Status	Precision	Recall	F1-Score	Support
Anxiety	0.80	0.75	0.77	1229
Bipolar	0.90	0.62	0.73	843
Depression	0.67	0.74	0.70	4585
Normal	0.83	0.95	0.89	4896
Personality Dis.	1.00	0.27	0.42	368
Stress	0.75	0.38	0.50	802
Suicidal	0.69	0.66	0.67	3190

TABLE III
PRECISION, RECALL, AND F1 SCORE FOR EVERY CLASS FOR THE BAG-OF-WORDS & SVM MODEL ON THE ENTIRE DATASET.

Status	Precision	Recall	F1-Score	Support
Anxiety	0.65	0.62	0.64	1179
Bipolar	0.57	0.56	0.57	889
Depression	0.65	0.67	0.66	4658
Normal	0.91	0.89	0.90	4892
Personality Dis.	0.48	0.45	0.47	353
Stress	0.52	0.54	0.53	806
Suicidal	0.57	0.58	0.58	3136

TABLE IV
PRECISION, RECALL, AND F1 SCORE FOR EVERY CLASS FOR THE BERT EMBEDDINGS & MLP MODEL ON THE ENTIRE DATASET.