

Enhancing Depression Detection in Social Media using Multimodal methods: A Scalable Approach for Improved Accuracy and Efficiency

Hang Sun, Junjie Li, Zihan Zhou, Xiaokang Ye, Pangyu Li, Yuheng Liu

{el20hs, el203jl, sc20zz2, mn20xy, mn20pl, sc20yl2}@leeds.ac.uk

1 Background

As the most prevalent mental disorder, depression is affecting over 300 million individuals worldwide (Russoniello, Fish and O'Brien, 2013). Moreover, evidence suggests that depression can be identified as a primary causal factor in more than two-thirds of suicides each year (Isometsä, 2014). Of concern is that since 2010, the incidence of depression has been increasing significantly with a non-negligible acceleration. Meanwhile, psychologists and medical doctors have found that the period seems to show a strong positive correlation with the start of the explosive diffusion of social media. Vanman, Baker and Tobin (2018) proposed that users' negative statements on social media may have serious psychological implications on the potentially depressed group, which rapidly dilates the depression rate. However, Internet communication is inevitable globally, where social media is utilized to help people share their thoughts, feelings, emotions, etc. While users are openly and freely discussing and responding, the output of depression is profoundly impacting the health of the depressed community. Simultaneously, the detection of depressive mood on user output can be regarded as the first stage of early prevention of depressive groups in the technological field. In summary, for computer scientists, the use of technology to detect and identify depressive statements of web users at an earlier stage plays an irreplaceable role in slowing down the development and spread of depression, which was also considered the basic motivation for this study.

Although this issue has received widespread academic attention, relatively few research projects have utilized data mining methods for early intervention in the disorder, so an exploration of more accurate methods for detecting depressed mood would be necessary. Several related research projects have been proposed, however, their practical effects do not seem to be fully applied to social media. For example, AudiB-ERT uses the multimodal nature of the human voice for screening depression (Toto et al., 2021), but users tend to rarely post vocal content on social media. Smys and Raj (2021) pro-

vided an early prediction of depressive patterns using a combination of support vector machines and the Naive Bayes algorithm, thus protecting them from the effects of mental illness and suicidal states. However, the method may ignore information about the emotional aspects of the text. Raymond's study (2021) analysed texts in social software that did not explicitly contain specific keywords such as "depression" or "diagnosis" to determine whether machine learning could be used effectively to detect signs of depression among social media users. Based on the complementary nature of images and text in social networks, it obviously ignores the emotional impact of images.

Data mining is a technique that can extract useful information from large-scale data. By analysing textual information in social networks, it can identify patients' emotions, behaviours and psychological states. Therefore, this research approach combines several aspects of data mining and proposes a framework to model different aspects of user text output using the Babbage model, RoBERTa model and ELECTRA model, and to detect sentiment on images using the CLIP model, and finally to unite the two models of graph and text in an ensemble manner. In addition, VADER sentiment scores are used to provide additional sentiment features.

When the program is applied, users, social platforms, healthcare professionals and government agencies will all benefit in some manner. Users will clearly gain health-related well-being by being prompted to seek treatment or minimizing the harm of depressive content. Moreover, healthcare professionals will have less pressure to work because the relevant tests have already been carried out earlier. Meanwhile, social networks will obtain a healthier communication environment and thus be protected from public censure. Finally, for government agencies, the application of the study could, to some extent, reduce the rate of depression in society.

In addition, this research will be declared AI and Data Science for Engineering, Health and Government (ASG) as part of the portfolio of EPSRC. The reasons can be identified as the follow-

ing two aspects. First, data mining is an important branch of data science. Second, the results of the study will have significant benefits for health and government.

Collectively, this study is a consolidation and remediation of past related studies worldwide. Primarily, our research approach combines various techniques that have performed well in past studies to model different aspects of text separately, for example, we used VADER sentiment scores to provide additional sentiment features. Also, we did not ignore the complementary nature of images and text and improved the accuracy of detection by exploiting image information.

2 Contribution to Knowledge

The research, which is at the intersection of the fields of natural language processing (NLP) and psychology, utilizes text and image information posted by users on social media as a dataset to determine whether the user has depressive symptoms using NLP sentiment analysis techniques. Therefore, it will expand the fields of NLP and psychology, and its algorithms are also involved in the fields of deep learning and artificial intelligence, where experts can morph and upgrade the algorithms and models to make them applicable to a wider range of fields.

This research provides a new technical tool in the field of psychology to detect depressed patients. In traditional psychology, depression screening usually requires questionnaires and face-to-face interviews, which are inefficient and inconvenient. More serious is that some patients will be reluctant to face a psychologist and have difficulty expressing themselves clearly inevitably, which will affect the detection accuracy (Van Lang, Ferdinand and Verhulst, 2007). This study offers a new approach to detecting depression more accurately, which is that after the assessment of users' social media postings through sentiment analysis, the doctor interviews users who are potentially at risk. This approach could provide researchers in the psychology field with a more efficient and accurate means of screening for depression, as well as help doctors to diagnose and treat depression more quickly. In addition, if psychology researchers were allowed access to content posted by depressed patients, it would facilitate researchers in analyzing patient personality and behavioural traits to further explore the pathogenesis and treatment of depression.

This research is also a new approach and idea in the NLP field in terms of sentiment detection. There are currently many challenges in this area, which can be more prominent when applied to social media. Common difficulties with sentiment analysis are associated with the polysemy and ambiguity of natural language and the complexity and subjectivity of human emotion. When it is applied to detecting social media content, sentiment analysis will face following severe difficulties. Firstly, social media texts are informal and contain quantities of colloquial terms, acronyms, emoticons, etc., which adds complexity to sentiment analysis. Secondly, the detection of social media texts requires a sharper capture of subtle sentiment signals to improve the accuracy of sentiment analysis than ordinary sentiment analysis models, as users often post very short messages to express their moods. In addition, a large amount of noise, such as the meaning of social media buzzwords, misspellings and grammatical errors, can also interfere with the accuracy of sentiment analysis results. Currently, experts in the field of natural language have used various algorithms in the sentiment analysis direction, which have provided good results, but they have their own limitations in providing high accuracy (Kumar and Gupta, 2021). As this project provides a valid and robust method for detecting signs of depression from social media information, it extends the scope of sentiment analysis applications and provides data support and methodological reference for other researchers who wish to perform depression detection.

In addition, the research will create value in the field of deep learning and artificial intelligence. We have reconstructed the textual dataset of social media, which provides a data reference for related research. Besides, as our model shows good results in predicting depression, this has implications for the optimization and improvement of sentiment analysis algorithms. Professionals can be able to adapt the algorithms from this study to other fields after modification and training, for example, the updated algorithms and models can be used in the evaluation of commercial products, the mindset of people reading the news, and the perception of political debates, etc.

To better interest every field, we will make our techniques more accessible to researchers and efforts to cooperate with relevant areas in the following ways. Firstly, we will make the

dataset and code publicly available and explain the code and training models used in the study, since an open-source project can facilitate participation and communication between academics and help others to better understand and reproduce our research. Moreover, we can likewise provide associations in psychology with texts that may present depressive moods, which can help them analyze patient characteristics and propose treatments.

3 Importance

In industry, depression detection on social networks has now gained wide visibility, which proves that there is a promising practical need and market prospect for this research. Many companies and organizations have begun to apply this technology to their products and services to improve user experience and mental health. For instance, Facebook has begun using automated algorithms to detect potentially depressed individuals and provide support and assistance accordingly (Fonseka, Bhat and Kennedy, 2019).

In the policy arena, it has also received attention and support from government and regulatory agencies, suggesting that the research does not violate existing privacy policies. The government is encouraging social networking platforms to take measures to prevent and deal with depression-related incidents by developing relevant policies and regulations. In addition, related research and development are being supported by the government to promote the development of depression detection technologies on social networks.

The country has played a pioneering role in the field of data mining from a relatively early stage and has now established an international leadership position. The detection of depression in social networks will greatly enhance our influence and competitive advantage in this domain, including the right to legislate and regulate international social platforms. More broadly, this research will contribute to the healthy evolution of our strong natural language processing technology, an important factor being the highly practical nature of the application of the project.

The country is also an international leader in depression control. The EPSRC have made significant and sustained investments in this area over many years, and consequently, the rate of depression in our country is chronically low using advanced engineering techniques. In combi-

nation, the research may lead to joint progress of different disciplines, i.e., the development of computer science leads to shared progress of psychology and medicine.

Research teams in various countries are now interested in this technology, and many countries globally (including the United States, Japan, France, etc.) have initiated major projects in the last decade. Therefore, ongoing investment in this technology is necessary not only to better control the incidence of depression but also to enable our country to remain at the forefront. In addition, this area will also be potentially beneficial for our digital economy. By eliminating the proliferating effect of social platforms on depression, technology companies will be shielded from public censure and reap the benefits of a larger user base, resulting in a larger online market.

Standing in the dimension of the globe, this project can reduce depression mood in the online environment and give an early indication of the output of depression, all of which show its significance to the global depression community.

4 Research Hypothesis And Objectives

4.1 Hypothesis

As mentioned earlier, although the detection of depressive tendencies in social networks has attracted both academic and industrial attention, the accuracy of current detection procedures applied to various social applications remains unsatisfactory. We believe that the reasons can be traced to the lack of updated use of the recent explosion of artificial intelligence technologies and the poor use of multimodal information.

One of the proposed and widely accepted research approaches is the use of BERT, a natural language model based on the transformer architecture, to model different aspects of the user's output to infer if the user is emotionally distressed.

However, BERT is already a technology proposed in 2018 and some iterations with better results have emerged in the industry, with well-known models such as OpenAI's latest release GPT-4 and RoBERTa proposed by Facebook in 2019. Recently, natural language models have come to a period of rapid development, and after experiments, we found that the fine-tuned Babbage model has been able to emotion detection to a high level of accuracy. This fine-tuning is the first part of our modelling for text. Meanwhile,

considering that the arithmetic power required by the Babbage model may be too large, we decided to use RoBERTa and ELECTRA models in the second part, and then fine-tune them with Vadder Score scores respectively, and finally combine the two models using an ensemble. In summary, our text depression recognition study is to train and experiment with the above two parts separately, and finally compare them and list the advantages and disadvantages. The specific choice is left to industrial B-side customers.

In addition, along with the version change of social networks, most of the user output has now been presented in the form of matching graphics and text. The theory, although proven to be excellent in its ability to recognize textual sentiment, does not seem to be well applied in contexts where graphics are combined with text. Here, we can give a more extreme example. In China, there was a widely spread internet stunt, "It's not really worthy to live". The text came from a Chinese celebrity's speech on a TV show and became a popular text among netizens for some time, but actually, the phrase was more of a playful, flirtatious expression of mind. Therefore, users often post this text with funny pictures. However, it is obvious that the current text depression recognition model will identify it as severe depression, which will cause some unnecessary misunderstanding as well as users' resentment. For the above reasons, we propose to use CLIP, a large model of Zero-shot pictures released by open AI in 2021 to detect the emotion of the pictures, and then ensemble the detection results of the text with those of the pictures and finally output the modal fusion with confidence.

In summary, the structure of our model can be roughly expressed as Figure 1.

4.2 Timeliness and novelty

Babbage is one of the latest natural language processing models based on GPT-3 developed by OpenAI, with up to 1.3 billion parameters. And CLIP is a large-scale neural network model developed by OpenAI in 2021 that has been pre-trained with over 400 million image-text pairs. To the best of our knowledge, both technologies have not yet been applied in the direction of depressive mood detection due to their recent release, which greatly demonstrates the timeliness of our project. We believe that with the support of the brand-new techniques, our research will bring a major breakthrough to depression detection,

and even to the whole field of sentiment analysis.

Our project combines multiple advanced models and techniques, including Babbage, RoBERTa, ELECTRA, Vadder Score, and CLIP, to achieve a high degree of accuracy in detecting depression in text and images. Then, the use of ensemble methods to combine the results of these models is also a novel approach that has been shown to improve the performance of the models. Finally, the use of CLIP to detect depression in images is also a unique and innovative application of the model.

4.3 Objectives

The long-term goal of this study is to continuously improve the accuracy of detecting depressive tendencies in social networks using advanced natural language processing techniques with image recognition. In future research, if more advanced techniques or methods with better results emerge or are developed, we will also consider adopting them into our research protocol, rather than sticking to the currently tailored protocol. The purpose of this study is to actively promote the application of natural language processing techniques in public health with a focus on depressed mood detection, while potentially enhancing the effectiveness of research in the field of sentiment analysis.

In particular, this project has the following subgoals to measure the results of our work.

- **Accuracy, precision, recall, and F1 score** of text / image / ensembled depressiveness detection in an acceptable range (specific values).

- **Ethical considerations met:** Evaluate the potential ethical implications of the system to ensure that it does not violate user privacy or perpetuate bias against specific populations.

- **Scalability compliance:** Evaluate the scalability of the system to ensure that it can handle large volumes of data and users without compromising performance.

- **Cost-effectiveness compliance:** Evaluate the cost-effectiveness of the system and identify any areas where costs can be reduced without compromising performance.

5 Programme And Methodology

5.1 Scenario definition

Based on our business understanding and target scenario, we abstracted our multimodal social

media post scenario as:

$$S = [(i_1, t_1, l_1), (i_2, t_2, l_2), \dots, (i_n, t_n, l_n)] \quad (1)$$

Where I denote the set of images, T denotes the set of texts, and L denotes the set of possible labels representing the depression degree, With $i_j \in I$, $t_j \in T$, and $l_j \in L$ for $j \in 1, 2, \dots, n$. Thus, the challenge of our depression detection can be attributed to predicate the l_j for any input of i_j and t_j from the user of social media.

5.2 Methodology and work programme

The research methodology faithfully adheres to the CRISP-DM standard, and we have incorporated an ample number of control groups. Under business understanding, we selected multimodal social media posts due to their practical significance. For data understanding, we analyzed public datasets and benchmarks but found a lack of suitable multimodal ones. Therefore, we chose a public dataset for detecting depression from posts as a text-based sentiment classifier benchmark. We annotated 96 images (32 per sentiment) as our initial image dataset, to be scaled in the next iteration.

To ensure a valid contrast experiment, we designed a sample-map mechanism to test our multimodal model for the first 75 images. This method employs three classification tag sets, L_1 , L_2 , and L_3 , a text dataset T , and an image dataset I to generate a multimodal test set D . Each example in the test set is a triplet (l, t_i, j) that represents a tag l , a text t_i matching the tag, and an image j matching the tag. Specifically, we first select an image j matching the tag from I , then select the text t_i matching the tag from T , and finally form a triplet (l, t_i, j) with the same tag l to add to the test set D . This method can be succinctly expressed as:

$$D = (l, t_i, j) \mid j \sim I, t_i \sim T, \text{ where } l \in t_i \text{ and } j \in l \quad (2)$$

where D is the test set, l is the label, t_i is the text of the matching tag, j is the image of the matching tag, T is the text dataset, I is the image dataset, and L_1 , L_2 , and L_3 are the three classification tag sets. This generated dataset can be used as the test set for our contrastive group and can be scaled up to make the result more valid in the second iteration. Also, To evaluate the performance of our model in some real-world scenarios with practical significance and challenges, such as sarcasm, irony, and so

on, we manually searched 21 tweets and images of different emotions from Twitter and added them to our test set.

For data preparation, we processed text and image inputs separately. Specifically, the depression dataset was converted into prompt/completion format for chatGPT finetuning using OpenAI's tool. For Bert, RoBERTa(Liu et al.,2019), and ELECTRA(Clark et al.,2020) models, Huggingface's autotokenizer formatted the input text.Nltk's Sentiment Intensity Analyzer was used to convert text into quads for Vader emotion scores.Pillow was used to convert images uniformly to RGB for CLIP's image input.

We propose a confidence-based modality fusion method that leverages both image input and text input information. The image input is processed using the pre-trained CLIP model (Radford et al., 2021), and a simple prompt [*not depression, moderate depression, severe depression*] is designed for the zero-shot CLIP model. The image, along with the prompt, is then inputted into the model to obtain a triplet of sentiment probabilities.

For the text-based input, we first obtain the nltk-VADER(Hutto et al.,2014) score based on the source text, which yields a tetrad of sentiment probabilities. Additionally, we tokenize the source text into word-based input for BERT-based models. Two fine-tuned models, RoBERTa and ELECTRA, are utilized to process the input and generate prediction results. The results are then encoded with a 128-dimension VAD embedding using MLP layers and an activation function. Finally, the output is aligned and ensemble(Ganaie et al.,2022), resulting in a triplet of sentiment probabilities.

We employ a decision-level modality fusion method based on confidence according to the two outputs: Let I denote the image-based modality and T denote the text-based modality. Let $O_I \in \mathbb{R}^3$ and $O_T \in \mathbb{R}^3$ represent the output arrays of the two models, respectively. The fusion method calculates the confidence of the image-based and text-based models by taking the maximum value of their respective output arrays:

$$C_I = \max(O_I), \quad C_T = \max(O_T) \quad (3)$$

The weights for each modality are determined by dividing its confidence value by the sum of both modalities' confidence values. This normalization ensures that the weights reflect the relative confi-

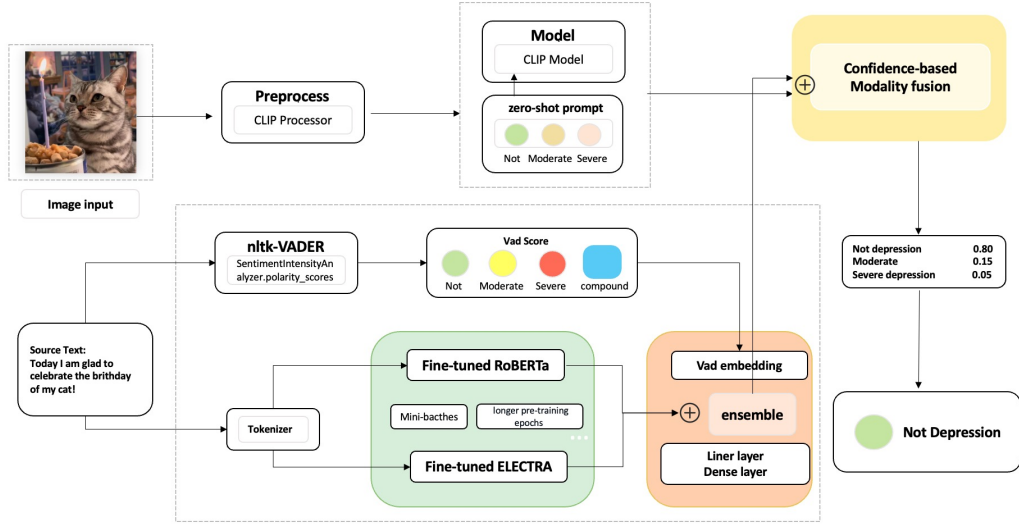


Figure 1: Our Multimodal Model Architecture

dence levels of each modality:

$$w_I = \frac{C_I}{C_I + C_T}, \quad w_T = \frac{C_T}{C_I + C_T} \quad (4)$$

Finally, the fusion method generates the final output by combining the image-based and text-based model outputs through element-wise multiplication and addition using the calculated weights:

$$O_F = w_I \odot O_I + w_T \odot O_T \quad (5)$$

where \odot denotes element-wise multiplication. The final output $O_F \in \mathbb{R}^3$ is a list of three sentiment probabilities, Thus we gain the final output of our model.

5.3 User Evaluation and proposed programme

For the evaluation of our multimodal classification method, we established several control groups utilizing a curated 96-image-text multimodal test dataset. The control groups included:

Method	Description
Baseline	An ensemble text-only classifier model which we developed, as described previously.
ChatGPT-babbage	ChatGPT-babbage, OpenAI's large-scale language model with 1.3 billion parameters, which was fine-tuned on our dataset.
Our full multimodal	Our full multimodal model which incorporates both visual and textual data.

We first conducted experiments on a set of 75 image-text test examples which were purposefully matched. We then tested the models on an additional 21 examples the challenge set comprised ambiguous real-world data to assess

generalizability as well as simulate the business-side user's real user experience.

The text-only model and ChatGPT-babbage serve as baselines to compare against our multimodal model.

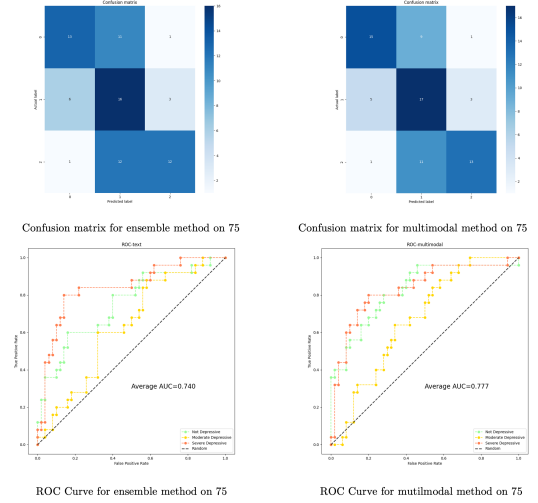


Figure 2: Confusion Matrix and ROC Curve for 2 methods on 75

	Fine-tuned LLM OpenAI's Babbage on 75 pair test set	RoBERTa&ELECTRA ensemble on 75 pair test set	Our Multimodal method on 75 pair test set
Accuracy	0.593	0.547	0.600
Precision	\	0.603	0.646
Recall	\	0.547	0.600
F1	0.608	0.554	0.606
AUC	\	0.740	0.777

Table 1: Performance of 3 Methods on 75 Pair Test Set

As demonstrated in the Table 1 and Figure 2, our proposed multimodal methodology achieved a notable performance on the 75-pairs dataset compared to the baseline text ensemble

approach, surpassing the text-based method in all criteria, rivalling even the capabilities of OpenAI’s large language model. As evidenced by the confusion matrix and ROC curve analyses, our proposed multi-modal approach leverages complementary information to substantively enhance model performance at a reduced cost. It achieves predictive abilities on par with those of the expensive LLM method, providing business users with a more affordable and capable alternative for monitoring emotional expressions on their platforms.

However, interrogation of the multimodal framework on the 21 ambiguous test cases revealed certain challenges. As tabulated in Table 2, the method failed to facilitate clarification of many ambiguous instances. That may attribute

	Text-based ensemble on 21 pair test set	Multimodal on 21 ambiguous test set
Accuracy	0.55	0.55
Precision	0.625	0.625
Recall	0.548	0.548
F1	0.562	0.563
AUC	0.723	0.759

Table 2: Performance of 2 Methods on 21 Pair Test Set

to fusion of modalities in our approach proceeded in an unsupervised, post-hoc manner at the decision level, rather than being integrated into model training. We did not leverage the multimodal signals when optimizing the model parameters when training, an avenue ripe for future improvement.

Upon establishing the validity and robustness of our approach through extensive experimentation, we will proceed to enhance both the breadth and depth of our model. Specifically, we intend to augment our dataset through strategic and comprehensive sampling in order to enhance the diversity and representativeness of our data. With an expanded, well-curated dataset, we will then be poised to retrain our model to maximize its performance on this dataset.

Concurrently, we aim to refine our model architecture to facilitate richer multi-modal interactions. Potential avenues for improvement include incorporating multiple modes of interaction between modalities, leveraging Transformer junctions and other architectures to enable information exchange across modalities, and aggregating the hidden representations from each modality to achieve fused modal understanding.

5.4 Contributions

Group members had specific responsibilities in the project, and we collaborated effectively. In our team, zihan, xiaokang and pangyu are responsible for the development and evaluation of algorithm model; junjie, hangsun and zihan are responsible for the background investigation, data collection, processing and writing; pangyu and xiaokang are responsible for the development of demo website. Yuheng was mainly responsible for fine-tuning of Babbage. We also help each other in different parts as partners. Stake-holders are primarily individual users, corporate users, media and regulatory bodies.

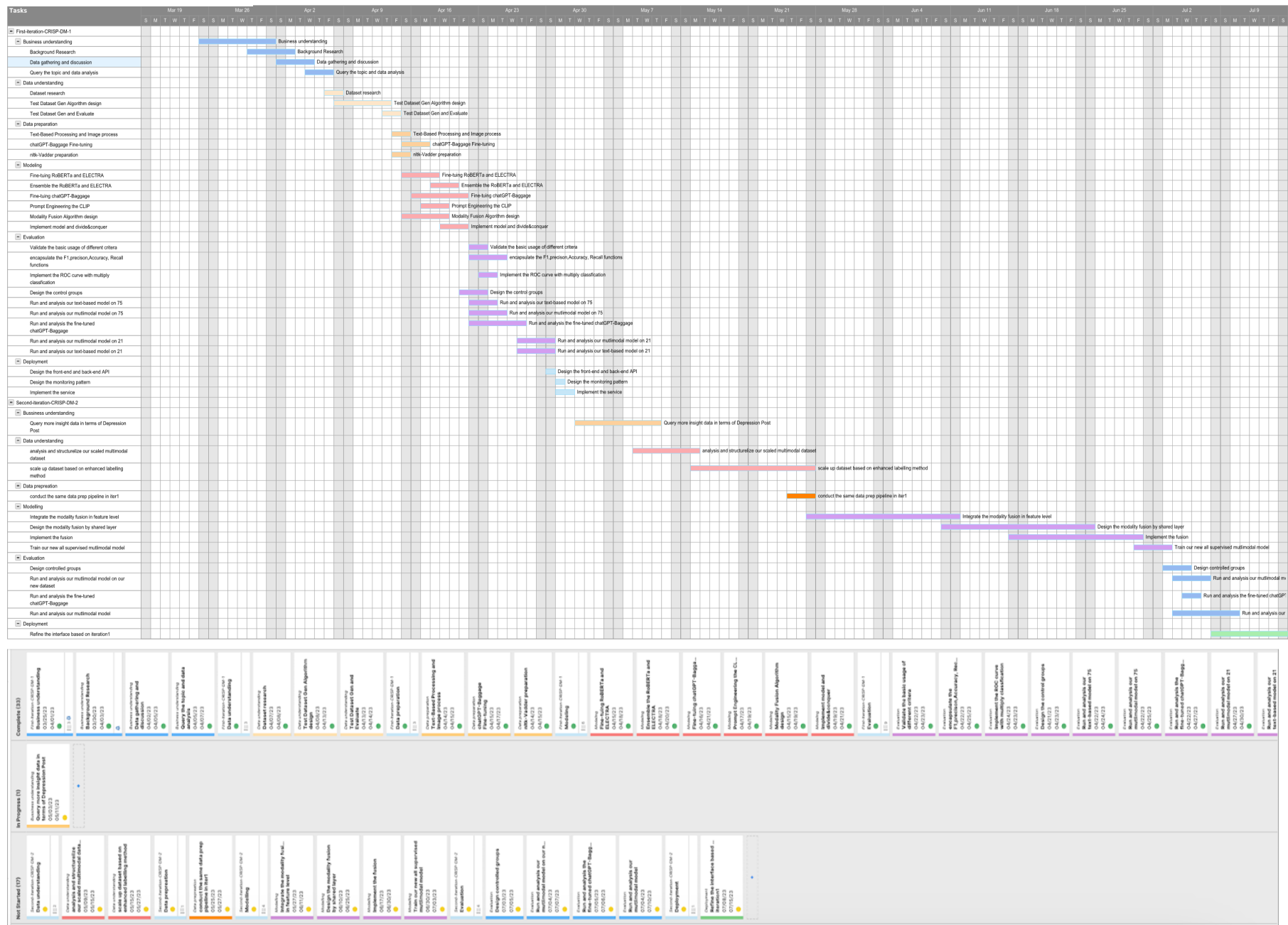
5.5 Project management and monitoring

Applying an iterative project management approach, this project followed the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology twice to gradually build an effective solution.

Initially, the CRISP-DM process was employed in an agile manner to fully explore the feasibility of the approach and identify opportunities for improvement. With an expanded dataset, the CRISP-DM methodology was repeated to further refine the solution. Specifically, a supervised fusion algorithm was trained to address more challenging tasks. By applying CRISP-DM iteratively with an increasing volume of data, the algorithm and problem understanding were progressively strengthened. In summary, following a dual iterative process enabled both rapid prototyping and gradual refinement of a solution to a complex problem.

Regarding project objectives and milestones, the first iteration aimed to comprehend the business scenario and propose a demonstration-level test dataset to evaluate our unsupervised modality fusion algorithm. The proposed dataset and algorithm showed its strengths and limitations, allowing optimal enhancement in the second iteration. The second iteration proposed an industry-level multimodal dataset and supervised modality fusion algorithm to address first iteration issues. Concluding two iterations, we gained a meaningful multimodal sentiment analysis dataset and mature multimodal algorithm. Each iteration achieved milestones finishing CRISP-DM stages: data understanding, modeling, evaluation; re-analysis and codebase implementation, as the work diagram detailed.

6.Work plan diagram



7 Reference

- Chiong, R. et al. (2021) "A textual-based featuring approach for depression detection using machine learning classifiers and social media texts," *Computers in Biology and Medicine*, 135, p. 104499. Available at: <https://doi.org/10.1016/j.combiomed.2021.104499>.
- Clark K. et al. (2020) ELECTRA: PRE-TRAINING TEXT ENCODERS AS DISCRIMINATORS RATHER THAN GENERATORS[C]. Available at: <https://openreview.net/pdf?id=r1xMH1BtvB>
- Fonseka, T.M., Bhat, V. and Kennedy, S.H. (2019) "The utility of artificial intelligence in suicide risk prediction and the management of suicidal behaviors," *Australian and New Zealand Journal of Psychiatry*, 53(10), pp. 954–964. Available at: <https://doi.org/10.1177/0004867419864428>.
- Ganaie, M.A. et al. (2022) "Ensemble deep learning: A review"[J].arXiv preprint arXiv:2104.02395v3. Available at: <https://arxiv.org/pdf/2104.02395.pdf>
- Hutto, C.J. Gilbert, E.E. (2014) VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.
- Isometsä, E. (2014) "Suicidal Behaviour in Mood Disorders—Who, When, and Why?," *The Canadian Journal of Psychiatry*, 59(3), pp. 120–130. Available at: <https://doi.org/10.1177/070674371405900303>.
- Kumar, A. and Gupta, D. (2021) "Sentiment Analysis as a Restricted NLP Problem," in *IGI Global eBooks*. IGI Global, pp. 65–96. Available at: <https://doi.org/10.4018/978-1-7998-4240-8.ch004>.
- Liu, Y. et al. (2019) "RoBERTa: A Robustly Optimized BERT Pretraining Approach"[J].arXiv preprint arXiv:1907.11692v1. Available at: <https://arxiv.org/pdf/1907.11692.pdf>
- Radford, A. et al. (2021) "Learning Transferable Visual Models From Natural Language Supervision"[J].arXiv preprint arXiv:2103.00020v1. Available at: <https://arxiv.org/pdf/2103.00020.pdf>
- Russoniello, C.V., Fish, M.T. and O'Brien, K.D. (2013) "The Efficacy of Casual Videogame Play in Reducing Clinical Depression: A Randomized Controlled Study," *Games for Health Journal*, 2(6), pp. 341–346. Available at: <https://doi.org/10.1089/g4h.2013.0010>.
- Smys, S. and Raj, J.S. (2021) "Analysis of Deep Learning Techniques for Early Detection of Depression on Social Media Network - A Comparative Study," *Journal of Trends in Computer Science and Smart Technology*, 3(1), pp. 24–39. Available at: <https://doi.org/10.36548/jtcsst.2021.1.003>.
- Toto, E., Tlachac, M. and Rundensteiner, E.A. (2021) AudiBERT, Conference on Information and Knowledge Management. Available at: <https://doi.org/10.1145/3459637.3481895>.
- Van Lang, N.D.J., Ferdinand, R.F. and Verhulst, F.C. (2007) "Predictors of future depression in early and late adolescence," *Journal of Affective Disorders*, 97(1–3), pp. 137–144. Available at: <https://doi.org/10.1016/j.jad.2006.06.007>.
- Vanman, E.J., Baker, R. and Tobin, S.J. (2018) "The burden of online friends: The effects of giving up Facebook on stress and well-being," *Journal of Social Psychology*, 158(4), pp. 496–508. Available at: <https://doi.org/10.1080/00224545.2018.1453467>.

8 Appendix

8.1 Demo

The demo model has been trained and encapsulated. The video's URL of running this demo is https://leeds365-my.sharepoint.com/:video/personal/el20hs_leeds_ac_uk/ES1SNu_SdBxGmi7MKNjPJooBaCk2H8JVz6WImJ6uUhKaTQ?e=zYcjJO.

The URL of code is <https://github.com/ProsperousYe/WALL-E-Detector.git> (Front-end)

<https://github.com/wall-e-DM/Multimodal-Depression-Detector.git> (Back-end).

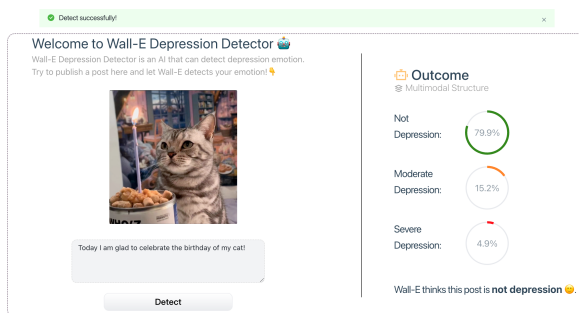


Figure 3: Our Demo Website

8.2 ChatGPT and GPT-4

ChatGPT has become the most widely discussed technology in the AI community since 2023, with its ability to answer questions, provide information, generate text and code, natural language processing and translation, and often achieve a very satisfactory result. During the writing of this report and the completion of the experiments and demos, the technique has been used in various ways, particularly in the sections on synthesising text and checking code for bugs. Due to the space limitations of the report, we will give some examples below.

8.2.1 Example 1

Adequate background knowledge of the relevant field is undoubtedly a prerequisite for conducting research. In order to understand text-based techniques for depressive mood detection, we have given the following hints about ChatGPT. (Due to space limitations ChatGPT's responses provide only the highlights)

Prompt: What would you suggest if I plan to learn some academic background on text-based

depression detection techniques using natural language processing techniques?

Completion: Learning the basics of natural language processing, reading relevant literature, exploring sentiment analysis, understanding machine learning and deep learning techniques, researching existing text-based depression detection techniques, attending relevant workshops and conferences, etc.

Prompt: You mentioned reading relevant literature, do you have any suggestions for keywords, search engines or websites etc.?

Completion: Keywords: depression detection, sentiment analysis, text-based sentiment analysis, natural language processing, machine learning, deep learning, medical natural language processing, social media text analysis, etc. Academic search engines: Google Scholar, Microsoft Academic, Semantic Scholar, Scopus, etc. Academic databases: ACM Digital Library, IEEE Xplore, SpringerLink, ScienceDirect, etc. Academic journals and conferences: Journal of Medical Internet Research, Computers in Human Behavior, ACM Transactions on Interactive Intelligent Systems, IEEE Transactions on Affective Computing, etc. Communities and forums: ResearchGate, Academia.edu, arXiv, etc.

So far, we have been given access to how to gain more background knowledge, which has a positive effect on completing the sections on the context of the report, contribution to knowledge and importance.

8.2.2 Example 2

The EPSRC requested that we elaborate on who might benefit from our research and how. However, as our team members were unfamiliar with the workings of some of the relevant industries, our original report only mentioned benefits to users, social platforms and government agencies. After iterating with ChatGPT using hint engineering, the end result broadened and hinted at relevant ideas for us, even though the answers were not quite what we had hoped for all at once.

Prompt: who might benefit from text-based depression detection technology?

Completion: People with depression, healthcare workers, counsellors and users of social media platforms.

Prompt: Can you describe in detail how this technology could benefit healthcare workers?

Completion: Text-based depression detection technology can help healthcare workers diagnose depression more quickly and accurately and provide better treatment and support to patients.

However, the answer seems to misunderstand the scenario in which our technology is used, i.e. our technology is not used in the real medical field, but to alert users in advance on social networking platforms. We are therefore correcting this answer.

Prompt: You seem to have misunderstood that our technology will not be used in the real medical field but on social networking platforms, and therefore it will not help healthcare workers to diagnose depression more quickly and accurately in patients.

Completion: When the technology is applied, potential patients will be alerted in advance, thus stopping the exacerbation of their condition, which will reduce the work pressure of healthcare workers.

In summary, the results of ChatGPT for healthcare workers were integrated and adopted into our report.

8.2.3 Example 3

In the Research Hypothesis and Objectives section, we were asked to describe why our report was timely and novel. After listing the ideas we could think of, we decided to use ChatGPT to give complementary suggestions.

We described in detail the methodology of our study and the timeliness and novelty we had in mind and concluded by asking: Please suggest more novelties and timeliness of our study.

Completion: In addition to the points you mentioned, the use of CLIP to detect depression in images is a unique and innovative application of this model.

Ultimately, that answer was also adopted as a point of novelty for us.

8.2.4 Example 4

In our research approach, the text-based detection model and the image-based detection model need to be encapsulated by ensemble to work together and thus identify the depressive tendencies in the user output. However, in the first round of comparison experiments, we found that the encapsulated model was not as accurate as, or even better than, the text-based model alone. Given that GPT-4 clearly improves the ability to understand and write code, we used GPT-4 to detect possible problems with our code.

We provide detailed code flow descriptions and code to ChatGPT and talk about the problems we face.

GPT-4 proposes that the VAD Score is not introduced in the ensemble module, but in a separate text detection model, which may have contributed to the lack of accuracy of our encapsulation model.

With its prompt, we added the code

```
# get vad score
_vad_score = _get_vad_score(text[0])
vad_score = torch.unsqueeze(_vad_score, 0)
```

Thus the Vadder Score is introduced correctly.

8.3 Google Scholar

Google Scholar is a search engine focused on scholarly literature, including articles, papers, books, conference papers and other scholarly publications. An example of how Google Scholar has contributed to our research is as follows.

Key words: Detection of Depression on Social Media Network

Search results: Analysis of Deep Learning Techniques for Early Detection of Depression on Social Media Network - A Comparative Study

After reading its abstract, we learned that they used a combination of support vector machines and a plain Bayesian algorithm to provide early prediction of depression patterns. I have therefore read the paper in detail and analysed the strengths and weaknesses of their approach in order to contribute to our research methodology.

8.4 Grammarly

We use Grammarly, which uses natural language processing and machine learning techniques to check, correct, and improve the quality of English writing, to improve our proposal writing quality before forming our formal proposal. Firstly before modifying our report, Grammarly asks us to specify what we expect to achieve, which has four components: audience, formality, domain, and intent. As this was a standard proposal to the EPSRC, we would set the knowledgeable audience, a formal writing style to regulate Grammarly to allow it to advise us on our article. After specifying our objectives, Grammarly can automatically check and correct spelling and grammatical mistakes, helping us write more accurate and fluent text. In addition, Grammarly can provide contextual suggestions for vocabulary and sentence patterns to make our writing more expressive and varied. Grammarly can provide tone suggestions based on our proposed goals, such as writing style and target audience, to make our text more relevant to our requirements. For example, sentence I was written without Grammarly, and Grammarly automatically detected the use of 'detecting' as a grammatical error and that 'contents' should be used in the singular in the current context. In addition, it also made suggestions for our style of writing, such as unclear sentences. Finally, we changed sentence II to sentence I. However, Grammarly needs some help with terminology and judgment. For example, it did not perform as well as the other sections when suggesting changes to the 'Programme And Methodology' section, as Grammarly is mainly used for general English writing and can produce incorrect detection and correction results for domain-specific terminology. Therefore, on crucial texts, Grammarly still requires human editors to ensure the accuracy and fluency of the text.

I. Seriously, when it is used to detecting contents in social media, sentiment analysis will face following difficulties.

II. When it is applied to detecting social media content, sentiment analysis will face following severe difficulties.

8.5 CRISP-DM

In creating the demo, we used the CRISP-DM process model and the project can be divided into the following phases:

Business understanding: In this phase, our team first defined the business problem and project goals. For example, the team wanted to develop models that could accurately detect the level of depression of social network users based on their text and image posts, and that would benefit a number of stakeholders, etc.

Data understanding: In this phase, our team collected and explored datasets related to the text and image posts of social network users. The team also analysed the data, understood the characteristics of the data, identified data quality issues and determined the feasibility of the project.

Data preparation: In this phase, the team cleaned, transformed and pre-processed the data in preparation for analysis.

Modelling: In this stage, the team developed two separate models using RoBERTa and ELECTRA. Each model should be fine-tuned using VAD Score scores. The team should also use the CLIP model to detect the level of depression in the images. The output of each model should be combined using ensemble techniques to enable graphically combined depression detection.

Evaluation: In this phase, the team evaluates the performance of the text and image models by measuring accuracy, precision, recall and F1 scores. The team also evaluates the performance of the integrated models by comparing their performance with that of the individual models.

Deployment: During this phase, the team deployed the final model to the web environment. The team developed a user interface that allowed social network users to enter their text and images and obtain depression-level predictions.

By following the CRISP-DM process model, the team ensured that each phase of the project was approached in a structured and systematic way, thus providing social network users with a more effective and efficient depression detection system.

8.6 Nltk VADER

NLTK's VADER is a prevalent sentiment analysis toolkit in natural language processing. As delineated in section 5.3, we employ it to quantify the affective valence of the utterance.