PyschExtract: Planning and Design Document

CM3070: Final Project
BSc Computer Science
University of London

CONTENTS

OVERVIEW

Mental-health documentation often relies on narrative notes, intake forms, and other unstructured text assessments that resist rigid structuring. Although these materials contain rich contextual and psychosocial information, much of this content is never systematically analysed. The manual coding of underlying insights (such as emotions, behavioural cues, and linguistic markers) is labour-intensive, inconsistent, and prone to subjective interpretation [1]. This makes large-scale analysis difficult and leaves valuable insights underutilised [2].

PsychExtract addresses this gap by automatically transforming mixed-format documentation (typed notes, scanned worksheets) into structured psychological insights. The system uses secure and lightweight pre-trained tools (such as Optical Character Recognition, semantic encoders, and Text-To-Speech) to extract emotional cues, linguistic patterns, and recurring themes. Real-world mental-health documentation varies widely in quality, therefore, each model within the pipeline is comparatively evaluated (Tesseract vs EasyOCR, RoBERTa vs DistilBERT, KeyBERT vs rule-based linguistic models, Coqui vs pyttsx3) to determine the most suitable and explainable approach.

Automating this early-stage extraction reduces documentation burden and supports consistent insight extraction, enabling mental-health professionals to focus on clinical reasoning rather than transcription or manual coding. Prior work demonstrates that automated extraction and summarisation meaningfully support mental-health monitoring and reduce administrative load [3, 4].

To coordinate these varying components, PsychExtract employs Template 4.1 (Orchestrating AI Models to Achieve a Goal). Mental-health documentation is inherently multimodal, in that no single model performs Optical Character Recognition (OCR), transformer-based Natural Language Processing (NLP), linguistic analysis, and Text-to-Speech (TTS) simultaneously. This modular orchestration approach therefore enables independent benchmarking of each pre-trained component, explainability through visible intermediate outputs, maintainability and future extensibility, and human-in-the-loop review (allowing users to inspect and edit generated insights).

This structure ensures the system remains transparent, interpretable, and user-centred. It also supports reproducibility, controlled experimentation, and rigorous performance comparison. This is both critical for mental-health research and for the project's primary objective of evaluating model performance across the pipeline.

The next sections outline the domain requirements, intended users, design decisions, architecture, work plan, feasibility analysis, and evaluation strategy.

DOMAIN AND USERS

DOMAIN

PsychExtract operates within the domain of psychological text analysis, where language is subjective, emotionally nuanced, and shaped by context. Systems in this domain must balance clarity, interpretability, and accuracy. Clarity is essential because psychological language is often ambiguous and unclear outputs can distort users' understanding of emotional content [5]. Interpretability equally important due to practitioners and researchers consistently report preferring transparent models whose reasoning can be traced and scrutinised [6]. Accuracy remains important as it underpins the reliability of emotion classification, yet even widely used datasets such as GoEmotions demonstrate macro-F1 scores around 0.46, highlighting the inherent difficulty of affective text interpretation [7]. These constraints shape PsychExtract's design toward explainability, transparent processing stages, and systematic comparative evaluation.

USERS

Three primary user groups inform the system's requirements. These are psychotherapists, psychology students, and psychological researchers, each with distinct expectations.

Psychotherapists benefit from concise and transparent summaries that supplement (not replace) their judgment. Prior work shows that clinicians adopt AI tools more readily when uncertainty is visible and the model's reasoning is inspectable [8]. PsychExtract therefore prioritises explainability, editable intermediate results, and conservative output framing.

Psychology students require an accessible way to explore how linguistic patterns relate to emotional meaning. Interpretable model outputs and editable OCR text support their learning without requiring technical expertise [5].

Psychological researchers need reproducible, modular pipelines that expose preprocessing decisions, classification steps, and error analysis. The system's modular design supports methodological transparency and replicable experimentation, consistent with current expectations in explainable affective NLP [9].

Framing the system around these three groups ensures that PsychExtract remains interpretable, trustworthy, and aligned with real-world mental-health practice.

DESIGN

PsychExtract integrates OCR, transformer-based NLP, linguistic interpretability layers, and TTS to automate the transformation of raw notes into structured insights. Automating this transformation reduces administrative burden and frees mental-health professionals to focus on synthesis, interpretation, and clinical care—reflecting well-established benefits of semi-automated documentation systems [10, 11]. A comparative modelling approach is applied across all components.

OCR

OCR converts handwritten or scanned text into machine-readable form. Tesseract and EasyOCR evaluated comparatively. Tesseract is computationally lightweight and reliable for clean printed text [12, 13], while EasyOCR's deep-learning architecture handles noisy variable inputs more robustly [14]. Evaluating both determines which is better suited to real unstructured mental-health documentation.

EMOTION CLASSIFICATION

Emotion classification facilitates affective insight central to psychosocial interpretation [15]. Two transformer models are assessed: DistilBERT, which is approximately 40% smaller and 60% faster than BERT while retaining 97% of its language understanding capabilities [16], making it a strong candidate for lightweight deployment; and RoBERTa, which typically yields a higher accuracy due to optimised pretraining [17]. The systematic benchmarking of both models balances performance with computational feasibility.

INTERPRETABILITY LAYER

Linguistic interpretability accompanies extracted emotion predictions to offer transparent, psychologically meaningful explanations [18, 19]. KeyBERT offers embedding-driven keyword extraction that captures semantic relevance [20], while alternative linguistic metrics (TF-IDF weighing or LIWC-style markers) offer rule-based, clinically explainable intuitive signals [18, 19]. Comparing these methods allows evaluation of semantic against rule-driven interpretability.

TTS

TTS provides multimodal accessibility for users with reading, attentional, or cognitive challenges [21]. Coqui TTS provides high-quality, neural speech synthesis [22], whereas pyttsx3 offers a reliable offline alternative suitable for constrained environments [23]. This comparison supports accessibility-oriented design, consistent with evidence that TTS enhances comprehension and engagement for diverse users [21].

DESIGN PRINCIPLES

The system follows core clinical-AI design guidelines [24]. The following principles are emphasised across the development of PsychExtract: modularity for safe isolation and replacement of models, transparency through visible intermediate layers, human-in-the-loop editing of OCR and review of generated insights, and minimal user interface (UI) for the purpose of emphasising clarity and explainability.

A functional UI implemented in Streamlit enables immediate testing. Figma supports early exploration of richer interface designs. This enables iterative refinement consistent with best practices in digital mental-health tool development [25].

ARCHITECTURE AND STRUCTURE

PsychExtract uses a modular, linear pipeline in which each stage is independently testable and replaceable. This maximises interpretability, facilitates comparative evaluation, and maintains clear traceability across the workflow. Data progresses through five main stages: OCR, emotion classification, linguistic interpretation extraction, summary generation, and optional TTS.

Pipeline Stages

1. Input Upload: user provides handwritten or scanned document
2. OCR: Tesseract/EasyOCR produces preliminary text
3. Editable Text: user verifies and corrects OCR output.
4. Emotion Classification: DistilBERT/RoBERTa produce multi-label probabilities
5. Interpretability Layer: KeyBERT or linguistic metrics extract explanatory features
6. Summary Generation: concise psychological insight produced
7. Optional TTS: pyttsx3/Coqui converts summary to speech
8. User Feedback Capture: supports iterative refinement and future evaluation

User Flow Diagram

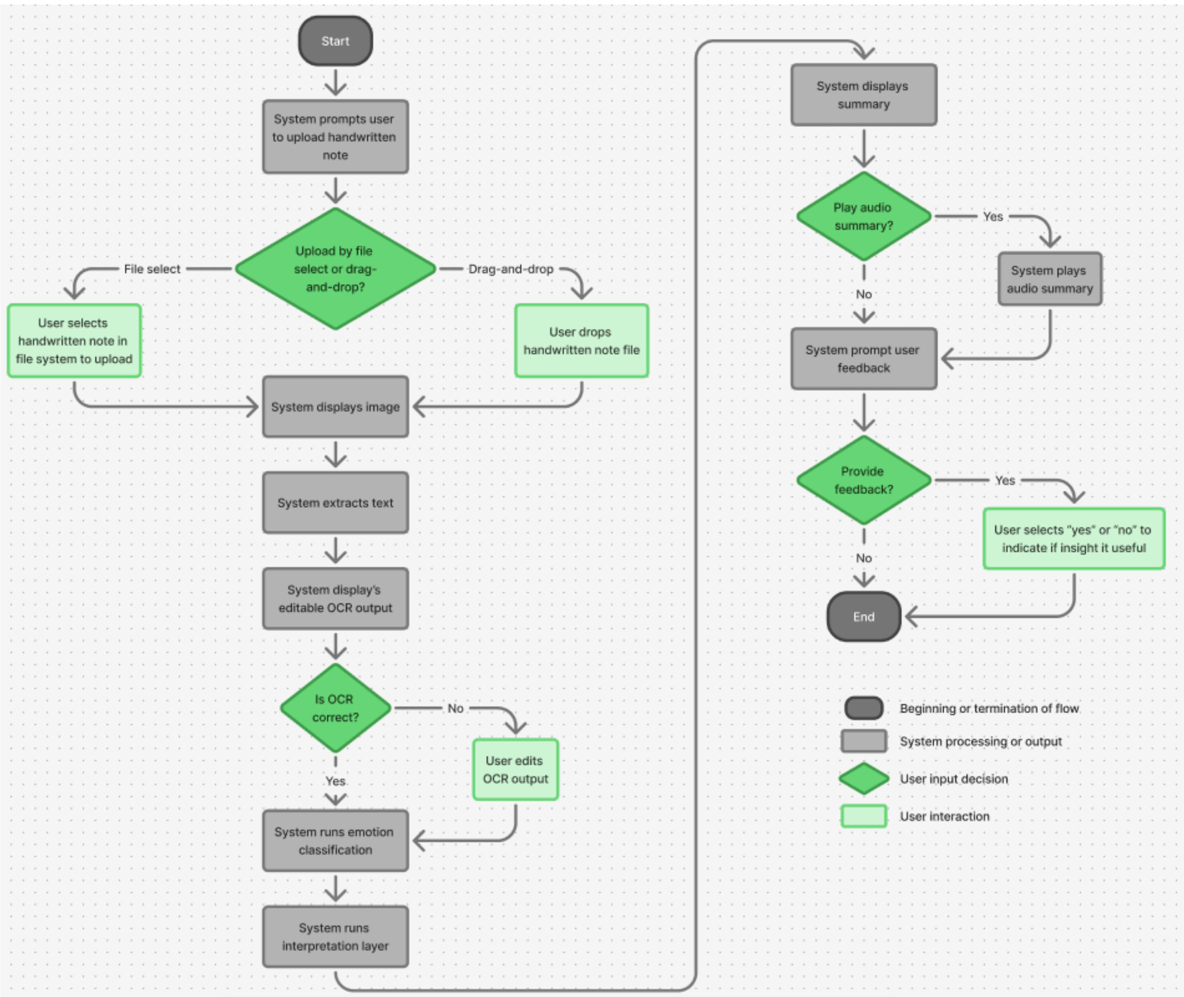

Figure 1. PsychExtract user flow

Early Prototype

A low-fidelity Streamlit UI demonstrates uploading, editing, and reviewing workflows

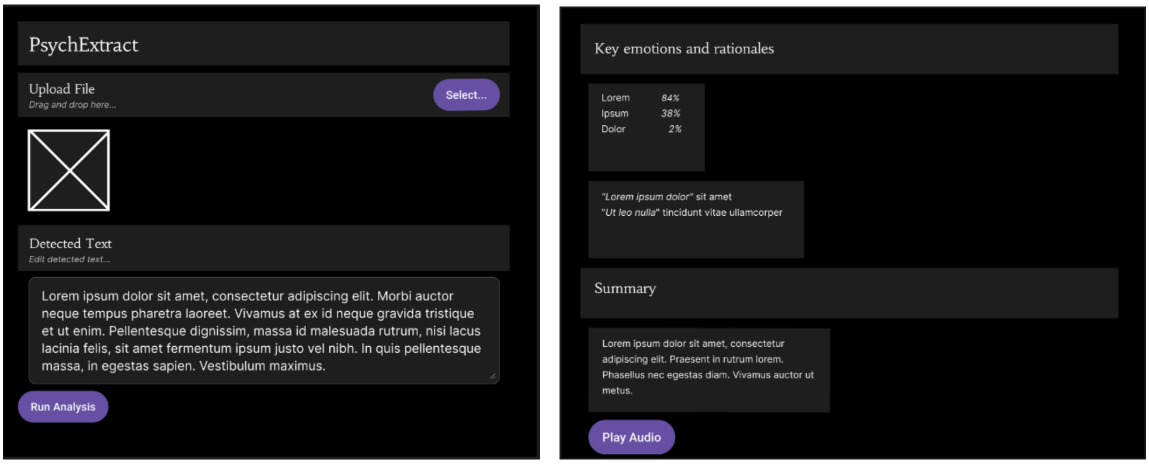

Figure 2. PsychExtract prototype interface

A more feature-rich annotator-style interface will be prototyped in Figma to explore alternatives without committing to full implementation.

Folder Structure

Folder	Purpose
ocr/	OCR wrappers for Tesseract and EasyOCR
nlp/	DistilBERT and RoBERTa models
interpret/	KeyBERT and linguistic metrics
summary/	Summarization and TTS components
ui/	Streamlit interface
tests/	Unit and integration tests

Table 1. PsychExtract folder structure

This modular organisation ensures traceability, flexible model swapping, and a clear separation of responsibilities between pipeline stages, facilitating efficient evaluation and iterative design.

Work Plan

The project runs from 8 December 2025 to 23 March 2026, following a structured iterative work plan divided across 4 months (see Figure 4). The workflow integrates continuous user testing and model evaluation.

Each major component (OCR, NLP, linguistic insights layer, TTS, and UI) follows a consistent cycle: a minimal working component is prototyped; it is evaluated through targeted tests (accuracy, robustness, performance, etc.); the component is subsequentially tested with users for clarity, usability, and interpretability; the component is then refined based on outcomes; finally it is integrated into the overall pipeline. Model comparison occurs within each phase, with insights feeding informing iterative refinement.

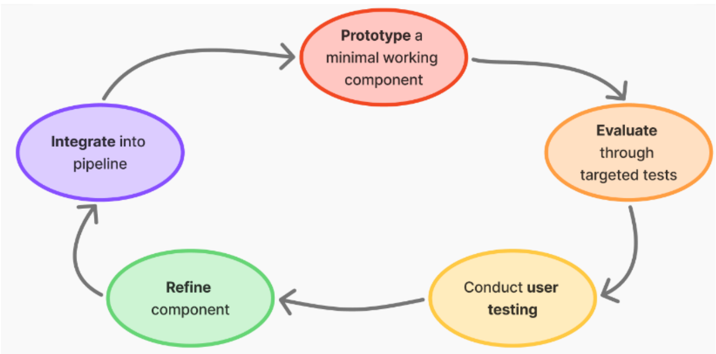

Figure 3. PsychExtract development cycle

Requirements and UI layouts are also tested early using low-fidelity prototypes, ensuring that design decisions reflect user needs before implementation. This continuous cycle of prototyping, evaluation, and refinement drives evidence-driven development and progressive system improvement from the initial concept to the final integrated system.

- Setup work
- Implementation work
- Evaluation work
- Reporting work

December 2025	Week 1	Week 2	Week 3	Week 4
Setup	**8–14 Dec** Prepare data Set up project environment Collect TTS resources	**15–21 Dec** Preprocess text Collect handwritten OCR samples		
Implementation and Fixes			**22-28 Dec** Build OCR prototype	
Testing and Evaluation	**8–14 Dec** Initial requirement gathering with user surveying	**15–21 Dec** Conduct short user test for early prototypes	**22-28 Dec** Evaluate OCR pipeline with different models	**29 Dec – 4 Jan** User testing on OCR models
Reporting			**22-28 Dec** Document OCR model results	**29 Dec – 4 Jan** Document OCR user testing results

January 2026	Week 1	Week 2	Week 3	Week 4
Setup				
Implementation and Fixes	**5-11 Jan** Build baseline NLP models	**12-18 Jan** Refine NLP feature set	**19-25 Jan** Implement rules for linguistic insights Integrate with NLP outputs	**26 Jan - 1 Feb** Finalize insight categories
Testing and Evaluation	**5-11 Jan** Test emotion classification	**12-18 Jan** Users review NLP outputs		**26 Jan - 1 Feb** Users rank insight categories
Reporting	**5-11 Jan** Document NLP model results	**12-18 Jan** Document NLP user testing results		**26 Jan - 1 Feb** Document insight model user testing results

February 2026	Week 1	Week 2	Week 3	Week 4
Setup				
Implementation and Fixes	**2-8 Feb** Integrate OCR, NLP, insights model Implement TTS		**16-22 Feb** Finalize UI structure	**23 Feb - 1 Mar** Build UI Integrate pipeline with UI
Testing and Evaluation	**2-8 Feb** Run end-to-end pipeline checks	**9-15 Feb** User testing on integrated prototype Evaluate TTS	**16-22 Feb** User testing on UI wireframes	**23 Feb - 1 Mar** Test UI with dataset
Reporting		**9-15 Feb** Document TTS model results Document TTS user testing results	**16-22 Feb** Document UI user testing	

March 2026	Week 1	Week 2	Week 3	Week 4
Setup				
Implementation and Fixes	**2-8 Mar** Integrate OCR, NLP, insights model Implement TTS			
Testing and Evaluation	**2-8 Mar** Evaluate system performance Collect final user satisfaction	**9-15 Mar** Polish system Conduct final debugging		
Reporting	**2-8 Mar** Document system evaluation results	**9-15 Mar** Bring together documented results into an academic report	**16-22 Mar** Polish final academic report	

Figure 4. PsychExtract December 2025 to March 2026 work plan

Feasibility and Contingency Plans

PsychExtract's design is realistic due to the modular architecture, lightweight model choices, and availability of both local and cloud-based compute. Transformer models used are compact enough for local execution, with Google Colab available as needed. The modular pipeline allows isolated debugging, clear performance benchmarking, and incremental development. The use of pre-trained models significantly reduces training cost and development time. Iterative testing embedded after each implementation stage enables early issue detection and prevents late-stage failures, supporting reliable progress toward final integration.

Contingencies strategies address component-level risks:
If both pre-selected OCR engines perform poorly on highly degraded scans, input may be restricted to digital PDFs, or stronger pre-processing can be applied. For extremely poor-quality documents, a text-only workflow can serve as a fallback.
If transformer-based NLP models exceed hardware limits, the TF-IDF logistic regression model (which is already part of the planned comparison) provides a fully functional alternative.
If interpretability outputs prove insufficiently clear to users, a lexicon-based psychological category system ensures stable and clinically interpretable fallback explanations.
For TTS, if neural synthesis causes latency or resource issues, the inclusion of pyttsx3 guarantees an offline, low-resource alternative.
For the UI, if clinical users are unavailable for testing and validation, psychology students or general usability participants will act as representative testers, as critical tasks focus on clarity and interaction, rather than psychological expertise.

Together, the comparative design, iterative testing schedule, and targeted contingencies ensure robust, resilient, and feasible development towards final system integration.

Evaluation Plans

The evaluation strategy is aligned with PsychExtract's objectives: assessing model performance, interpretability, and user experience across all pipeline stages. Evaluation is embedded throughout development and repeated after full integration to ensure component-level insights translate into coherent system behaviour.

OCR Evaluation

OCR evaluation focuses on accuracy, robustness, and contextual error patterns. Character-level accuracy and word-error rate serve as core quantitative metrics, capturing both fine-grained recognition quality and overall textual coherence. Quantitative outcomes are supplemented by structured error analysis that identifies recurrent failure types (such as letter-shape confusions, mis-segmentation, and spacing considerations) offering insight into which OCR engine is better suited to variable real-world mental-health documentation.

NLP Classification Evaluation

Emotion classifiers are assessed using precision, recall, and macro-F1, which measure the quality of multi-label predictions, reflecting the uneven class distributions typical of affective datasets. Confusion matrices reveal systematic patterns of misclassifications, while qualitative error analysis examines mislabelled samples to determine whether model limitations, linguistic ambiguity, or upstream OCR noise. This dual analysis supports transparency by explaining how errors occur and why.

Interpretability Layer Evaluation

The interpretability layer is assessed through both quantitative coherence measures and qualitative judgement from target users. Topic coherence scores such as UMass and c_v assess whether extracted key-phrases or clusters reflect meaningful semantic groupings [26]. Complementary readability indices, including Flesch–Kincaid [27] and lexical diversity metrics, quantify clarity and linguistic complexity in generated insights. User-testing oriented face-validity checks assess whether outputs appear plausible and psychologically aligned, ensuring interpretability remains grounded in real-world expectations. Integration checks verify that upstream errors do not propagate to distort downstream explanations.

TTS Evaluation

TTS performance considers both technical responsiveness and subjective usability. Latency (the time between text generation and audio output) indicates computational efficiency, while a small MOS-style naturalness rating [28] (where participants assess clarity on a scale of 1 to 5) captures perceived clarity and listening comfort.

Interaction and Usability Evaluation

Interaction and usability evaluation centres on transparency, trust, and ease of use. Cognitive walkthroughs examines whether new users can successfully upload documents, correct OCR results, understand extracted insights, and initiate TTS playback. A small number of users perform task-based usability tests (capturing behavioural observations, think-aloud reflections, and short post-task questionnaires) revealing friction points and guiding iterative refinement.

This evaluation plan ensures methodological rigour, supports comparative analysis across models, and maintains alignment with the system's aims of interpretability, transparency, and user-centred design.

References

[1] Theresa A. Koleck, Caitlin Dreisbach, Philip E. Bourne, *et al.* 2019. Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. *Journal of the American Medical Informatics Association* 24, 4 (Apr. 2019), 364-379. https://doi.org/10.1093/jamia/ocy173

[2] Sankha S. Mukherjee, Jiawei Yu, Yida Won, *et al.* 2020. Natural language processing-based quantification of the mental state of psychiatric patients. *Computational Psychiatry* 5 (Dec. 2020), 76-106. https://doi.org/10.1162/cpsy_a_00030

[3] Rosanne J. Turner, Femke Coenen, Femke Roelofs, *et al.* 2022. Information extraction from free text for aiding transdiagnostic psychiatry: constructing NLP pipelines tailored to clinicians' needs. *BMC Psychiatry* 22, Article 407 (Jun. 2022), 11 pages. https://doi.org/10.1186/s12888-022-04058-z

[4] Jessica P. Ridgway, Amo Uvin, Jessica Schmitt, *et al.* 2021. Natural language processing of clinical notes to identify mental illness and substance use among people living with HIV:

retrospective cohort study. *JMIR Medical Informatics* 9, 3, Article e23456 (Mar. 2021), 10 pages. https://doi.org/10.2196/23456

[5] James W. Pennebaker, Ryan L. Boyd, Kayla Jordan, *et al.* 2015. The Development and Psychometric Properties of LIWC2015. *University of Texas at Austin.* https://doi.org/10.13140/RG.2.2.23890.43205

[6] Finale Doshi-Velez and Been Kim. 2017. Towards a Rigorous Science of Interpretable Machine Learning. *arXiv preprint.* https://doi.org/10.48550/arxiv.1702.08608

[7] Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, *et al.* 2020. GoEmotions: A Dataset of Fine-Grained Emotions. In *Proceedings of ACL 2020* (Jul. 2020), 4040–4054. https://doi.org/10.18653/v1/2020.acl-main.372

[8] Maia Jacobs, Jeffrey He, Melanie Pradier, *et al.* 2021. Designing AI for Trust and Collaboration in Time-Constrained Medical Decisions: A Sociotechnical Lens. *In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21).* Association for Computing Machinery 659, 1-14. https://doi.org/10.1145/3411764.3445385

[9] Saif M. Mohammad. 2022. Ethics Sheet for Automatic Emotion Recognition and Sentiment Analysis. *Computational Linguistics* 48, 2 (Jun. 2022), 239-278. https://doi.org/10.1162/coli_a_00433

[10] Dror Ben-Zeev, Kristin E. Davis, Susan Kaiser, *et al.* 2013. Mobile technologies among people with serious mental illness: opportunities for future services. *Administration and Policy in Mental Health and Mental Health Services Research* 40, 4 (Jul. 2023), 340-343. https://doi.org/10.1007/s10488-012-0424-x

[11] Tait D. Shanafelt, Lotte N. Dyrbye, Christine Sinsky, *et al.* 2016. Relationship between clerical burden and characteristics of the electronic environment with physician burnout and professional satisfaction. *In Mayo Clinic Proceedings* 91, 7 (Jul. 2016), 836-848. https://doi.org/10.1016/j.mayocp.2016.05.007

[12] Ray W. Smith. 2007. An Overview of the Tesseract OCR Engine. *In Proceedings of ICDAR '07 Vol. 2* (Sep. 2007), 629–633. https://doi.org/10.1109/ICDAR.2007.4376991

[13] Chirag I. Patel, Atul Patel, and Dharmendra Patel. 2012. Optical Character Recognition by Open Source OCR Tool Tesseract: A Case Study. *International Journal of Computer Applications* 55, 10 (Oct. 2012), 50-56. https://doi.org/10.5120/8794-2784

[14] JaidedAI. 2024. EasyOCR: Ready-to-use OCR with 80+ supported languages and all popular writing scripts including: Latin, Chinese, Arabic, Devanagari, Cyrillic, etc. *GitHub.* Retrieved Nov. 25 2025 from https://github.com/JaidedAI/EasyOCR, *archived at* [https://web.archive.org/web/20251012081255/https://github.com/JaidedAI/EasyOCR]

[15] Leslie S. Greenberg and Antonio Pascual-Leone. 2006. Emotion in psychotherapy: A practice-friendly research review. *Journal of Clinical Psychology* 62, 5 (2007), 611–630. https://doi.org/10.1002/jclp.20252

[16] Victor Sanh, Lysandre Debut, Julien Chaumond, *et al.* 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint.* https://doi.org/10.48550/arXiv.1910.01108

[17] Liu Yinhan, Myle Ott, Naman Goyal, *et al.* 2019. Roberta: A robustly optimized BERT pretraining approach. *arXiv preprint*. https://doi.org/10.48550/arXiv.1907.11692

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, *et al.* 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* 1 (Jun. 2019), 4171-4186. https://doi.org/10.18653/v1/N19-1423

[19] Becky Inkster, Ross O'Brien, Emma Selby, *et al.* 2020. Digital Health Management During and Beyond the COVID-19 Pandemic: Opportunities, Barriers, and Recommendations. *JMIR mental health* 7, 7, Article e19246 (Jul. 2020), 5 pages. https://doi.org/10.2196/19246

[20] Maarten Grootendorst. 2025. KeyBERT. *GitHub*. Retrieved Nov. 26 2025 from https://github.com/MaartenGr/KeyBERT, *archived at* [https://web.archive.org/web/20250927082041/https://github.com/MaartenGr/KeyBERT]

[21] Mary Cece Young, Carrie Anna Courtad, Karen H Douglas, *et al.* 2018. The Effects of Text-to-Speech on Reading Outcomes for Secondary Students With Learning Disabilities. *Journal of Special Education* Technology 24, 2 (Jul. 2018), 80-91. https://doi.org/10.1177/0162643418786047

[22] Coqui.ai. 2025. Coqui.ai TTS. *TTS 0.22.0 documentation*. Retrieved Nov. 26 from https://docs.coqui.ai/en/latest, *archived at* [https://web.archive.org/web/20250812191341/https://docs.coqui.ai/en/latest]

[23] pyttsx3.readthedocs.io. 2025. pyttsx3 - Text-to-speech x-platform. *pyttsx3 2.6 documentation*. Retrieved Nov. 26 2025 from https://pyttsx3.readthedocs.io/en/latest, *archived at* [https://web.archive.org/web/20251010005655/https://pyttsx3.readthedocs.io/en/latest]

[24] Medicines and Healthcare Products Regulatory Agency (MHRA). 2025. Software and artificial intelligence (AI) as a medical device. *gov.uk*. Retrieved Nov 27. 2025 from https://www.gov.uk/government/publications/software-and-artificial-intelligence-ai-as-a-medical-device/software-and-artificial-intelligence-ai-as-a-medical-device, *archived at* [https://web.archive.org/web/20250815022341/https://www.gov.uk/government/publications/software-and-artificial-intelligence-ai-as-a-medical-device/software-and-artificial-intelligence-ai-as-a-medical-device]

[25] John Torous, Jennifer Nicholas, Mark E. Larsen, *et al.* 2018. Clinical review of user engagement with mental health smartphone apps: evidence, theory and improvements. *Evidence-based mental health* 21, 3 (Aug. 2018), 116-119. https://doi.org/10.1136/eb-2018-102891

[26] Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the Space of Topic Coherence Measures. *In Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM '15)* (Feb. 2015), 399-408. https://doi.org/10.1145/2684822.2685324

[27] Peter J. Kincaid, Robert P. Fishburne Jr., Richard L. Rogers, *et al.* 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. *Technical Report Research Branch Report, Millington, TN: Naval Technical Training*. Article RBR875, 51 pages. https://doi.org/10.21236/ADA006655

[28] ITU-T P.800. 1996. Methods for subjective determination of transmission quality. *International Telecommunication Union*. 37 pages.