

Conceptual Analysis of Psychological Terms Using APA Thesaurus: A Comparative Study of Term Extraction Methods

Florence Zhao

New York University
tz2556@nyu.edu

Yifei Zhao

New York University
yz9704@nyu.edu

Lixing Zong

New York University
lz3070@nyu.edu

Abstract

Understanding psychological terminology is fundamental to advancing research in both psychology and natural language processing. This paper presents a comprehensive study on domain-specific term extraction from psychology texts, comparing three approaches: TF-IDF, RAKE, and MentalBERT. We construct a manually annotated dataset comprising 52 psychology abstracts with 603 annotated terms, and evaluate extraction quality using multiple metrics including Precision, Recall, F1, and Precision@K. Our experiments demonstrate that MentalBERT achieves an F1 score of 0.250, nearly doubling the F1 of 0.134 achieved by the best baseline method (RAKE). Furthermore, we integrate the APA Thesaurus of Psychological Index Terms to conduct conceptual analysis, achieving a clustering purity of 0.692 and normalized mutual information of 0.127. Error analysis reveals that abstract theoretical concepts pose greater extraction challenges than concrete experimental terms, with performance varying significantly across psychological subfields. These findings highlight the importance of domain adaptation in terminology extraction and provide insights for future research in psychology-specific NLP applications.

1 Introduction

Psychological terminology presents unique challenges for natural language processing systems. Unlike domains such as medicine or finance where terminology tends to be more concrete and standardized, psychological language often involves abstract constructs with context-dependent meanings. Terms such as “self-efficacy,” “cognitive bias,” and “attachment style” carry rich theoretical implications that standard keyword extraction methods frequently fail to capture.

The motivation for this research stems from the observation that existing term extraction approaches, while successful in many scientific domains, struggle with the nuanced vocabulary of psychology. Statistical methods like TF-IDF excel at identifying frequently occurring terms but often miss theoretically important concepts that appear less frequently. Rule-based approaches such as RAKE can capture multi-word expressions but lack the semantic understanding necessary to distinguish domain-specific terminology from general academic language.

This study addresses these challenges through three main contributions. First, we develop a domain-specific terminology extraction system tailored for psychology texts, implementing and comparing TF-IDF, RAKE, and MentalBERT approaches. Second, we create a manually annotated dataset of psychological terminology from 52 academic abstracts, establishing a gold standard for evaluation. Third, we integrate the APA Thesaurus of Psychological Index Terms ([American Psychological Association, 2023](#)) to conduct conceptual analysis, evaluating how well extracted terms align with established psychological categories.

Our experimental results demonstrate that domain-adapted transformer models significantly outperform traditional baselines for this task. MentalBERT achieves an F1 score of 0.250, compared to 0.134 for RAKE and 0.120 for TF-IDF. Beyond quantitative improvements, our semantic clustering analysis reveals meaningful patterns in how different methods capture psychological concepts, with implications for both NLP methodology and psychological research applications.

2 Related Work

Research on automatic term extraction has evolved considerably over the past two decades, progress-

ing from purely statistical methods to sophisticated neural approaches. This section reviews key developments relevant to our work on psychology-specific term extraction.

2.1 Statistical Term Extraction

Early approaches to term extraction relied primarily on statistical measures of term importance. [Frantzi et al. \(2000\)](#) introduced the C-value/NC-value method, which combines linguistic pattern matching with statistical analysis to identify multi-word terms. This approach became a foundational benchmark for terminology extraction in technical domains. TF-IDF (Term Frequency-Inverse Document Frequency) remains widely used as a baseline method, offering computational efficiency and interpretability despite its inability to capture semantic relationships between terms.

[Rose et al. \(2010\)](#) proposed RAKE (Rapid Automatic Keyword Extraction), a rule-based algorithm that uses stopword boundaries and word co-occurrence to identify keyphrases. RAKE’s strength lies in its unsupervised nature and ability to capture multi-word expressions without requiring training data. However, like TF-IDF, it lacks semantic understanding and tends to favor longer phrases regardless of their domain relevance.

2.2 Neural and Embedding-based Methods

The advent of pre-trained language models has transformed keyword extraction. KeyBERT ([Groothedest et al., 2020](#)) leverages BERT embeddings to identify semantically representative keywords by computing cosine similarity between document and candidate phrase embeddings. This approach captures semantic relationships that statistical methods miss, though it may struggle with highly specialized domain vocabulary not well-represented in general-purpose pre-training corpora.

Domain-specific transformer models have shown promise for specialized text processing. Mental-BERT ([Ji et al., 2022](#)) was pre-trained on mental health text from Reddit and demonstrates improved performance on mental health classification tasks. Similarly, PsyBERT and other psychology-focused models have emerged to address the unique linguistic characteristics of psychological discourse.

2.3 Multi-word Expression Processing

Psychological terminology frequently involves multi-word expressions (MWEs) such as “cognitive behavioral therapy” or “social anxiety disor-

der.” [Savary et al. \(2017\)](#) established annotation frameworks for verbal MWEs in the PARSEME shared task, providing methodological foundations for multi-word term identification. Their work on cross-lingual MWE annotation informs our approach to psychology term annotation.

2.4 Domain-specific Term Extraction

[Luan et al. \(2018\)](#) presented a multi-task neural framework for extracting entities and relations from scientific text, demonstrating that contextual embeddings significantly improve extraction quality in specialized domains. Their work provides a foundation for our exploration of embedding-based term extraction in psychology.

Graph-based methods like TopicRank ([Bougouin et al., 2013](#)) offer an alternative approach by modeling document structure as a graph and ranking candidate keyphrases based on their centrality. While effective for general documents, these methods may require domain adaptation to handle the specialized vocabulary and conceptual structure of psychology texts.

Our work builds upon these foundations by systematically comparing traditional and neural approaches specifically for psychology terminology, integrating the APA Thesaurus for domain-grounded evaluation, and conducting detailed error analysis across psychological subfields.

3 Data

3.1 Data Collection

We collected psychology abstracts from multiple academic sources to ensure coverage across psychological subfields. Our corpus comprises 385 papers drawn from three primary sources: 186 papers from arXiv (covering computational approaches to psychology), 194 papers from psychology-specific collections including APA journals and ResearchGate, and 5 classic foundational papers in psychology theory.

The collected texts span five major psychological subfields: cognitive psychology, developmental psychology, social psychology, clinical psychology, and personality psychology. This diversity ensures that our evaluation captures the varying terminological characteristics across different areas of psychological research. All texts are academic in nature, using formal domain-specific vocabulary appropriate for evaluating term extraction systems.

Statistic	Value
Total Documents	52
Total Term Instances	603
Unique Terms	558
Avg Terms per Document	11.60
Avg Term Length (words)	2.12
Term Diversity	92.5%

Table 1: Summary statistics for the annotated development dataset used for evaluation.

3.2 Annotation Process

We developed a human-AI collaborative annotation procedure for identifying psychological terms. Given the scale of annotation required and resource constraints, we employed a large language model (gpt-5-mini) to generate initial term annotations based on predefined guidelines specifying that terms should represent psychological concepts, theoretical constructs, methodological approaches, or established phenomena. Team members then reviewed and verified the LLM-generated annotations, correcting obvious errors and ensuring consistency with domain conventions.

This approach trades off traditional inter-annotator agreement measurement for scalability. While we cannot report Cohen’s Kappa between independent human annotators, the LLM-assisted process offers reproducibility and consistency that purely manual annotation may lack. We acknowledge this as a limitation and discuss its implications in the Limitations section.

The annotation effort focused on a development set of 52 abstracts, yielding 603 annotated term instances with 558 unique terms. On average, each document contains 11.6 annotated terms, reflecting the terminology-dense nature of psychology abstracts. The distribution of term lengths shows that psychological terminology ranges from single words (e.g., “cognition”) to multi-word phrases (e.g., “cognitive behavioral therapy”), with an average term length of 2.12 words.

Table 1 summarizes the key statistics of our annotated dataset. The high term diversity (92.5%) indicates that psychological terminology exhibits considerable variety, with relatively few terms appearing across multiple documents.

3.3 APA Thesaurus Integration

To ground our evaluation in established psychological taxonomy, we integrated the APA Thesaurus of Psychological Index Terms (12th edition) ([American Psychological Association, 2023](#)). This thesaurus provides standardized conceptual categories including Emotion Processes, Cognitive Mechanisms, Social Psychology, Developmental Psychology, Clinical Psychology, Personality Psychology, Neuropsychology, and Methodology.

We implemented an automated mapping system that categorizes extracted terms into these APA categories based on lexical matching and keyword patterns. Terms that do not match any established category are labeled as “Uncategorized.” This integration enables us to evaluate not only whether extraction methods identify correct terms, but also whether they capture psychologically meaningful concepts that align with recognized theoretical frameworks.

3.4 Data Splits

Following standard practice in NLP evaluation, we partitioned our data into development and test sets. The development set of 52 annotated documents serves as our primary evaluation corpus for algorithm comparison and hyperparameter tuning. A separate test set is held out for final evaluation to prevent overfitting to development data. Note that our extraction methods (TF-IDF, RAKE, MentalBERT) do not require supervised training on labeled examples; MentalBERT uses pre-trained weights without task-specific fine-tuning.

4 Methodology

Our system implements three term extraction approaches, progressing from statistical baselines to neural methods. This section describes each approach and our evaluation framework.

4.1 Preprocessing

All input texts undergo standardized preprocessing before extraction. We apply tokenization and sentence segmentation using SpaCy, followed by lemmatization and part-of-speech tagging for linguistic normalization. Stopword removal uses an extended English stopword list augmented with psychology-specific stopwords (e.g., “study,” “results,” “experiment,” “analysis”) that appear frequently in academic writing but carry little domain-specific meaning.

4.2 TF-IDF Baseline

Our first baseline implements TF-IDF (Term Frequency-Inverse Document Frequency) extraction. For a term t in document d within corpus D , the TF-IDF score is:

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \cdot \log \frac{|D|}{\text{DF}(t)} \quad (1)$$

where $\text{DF}(t)$ is the number of documents containing term t .

We configure the TF-IDF vectorizer to consider n-grams ranging from unigrams to trigrams (`ngram_range=(1,3)`), enabling capture of multi-word psychological terms. A maximum document frequency threshold of 0.85 filters out overly common terms. For each document, we extract the top 10 terms ranked by TF-IDF score.

4.3 RAKE Baseline

RAKE (Rapid Automatic Keyword Extraction) serves as our second baseline. This algorithm identifies candidate phrases by using stopwords and phrase delimiters as boundaries, then scores candidates based on word co-occurrence patterns. For each candidate phrase, RAKE computes a score as the sum of word scores, where each word’s score is the ratio of its degree (co-occurrence frequency) to its frequency.

We configure RAKE to extract phrases of 1 to 3 words, matching the typical length of psychological terms observed in our annotation. The top 10 ranked phrases are selected as extracted terms for each document.

4.4 MentalBERT Extraction

Our primary extraction method leverages MentalBERT, a domain-adapted BERT model pre-trained on mental health text. We integrate MentalBERT with KeyBERT’s extraction framework to combine domain-specific embeddings with semantic similarity-based keyword selection.

The extraction process operates in two stages. First, we generate candidate terms by identifying noun phrases and n-grams from the input text. Second, we embed both the full document and candidate terms using MentalBERT, then rank candidates by their cosine similarity to the document embedding.

To improve extraction quality, we implement multiple extraction strategies. The first strategy applies Maximal Marginal Relevance (MMR) with

diversity parameter 0.4 to balance relevance and diversity among extracted terms. The second strategy uses Max Sum similarity to select terms that are maximally similar to the document while minimally similar to each other. We combine results from both strategies and apply post-processing filters.

Post-processing includes deduplication, length filtering (removing phrases longer than 3 words), and domain relevance filtering using the APA Thesaurus. Terms that match APA categories or contain psychology-related keywords are prioritized in the final ranked list.

4.5 Evaluation Metrics

We employ a comprehensive evaluation framework encompassing both extraction quality and conceptual relevance metrics.

Basic Metrics. Precision measures the fraction of extracted terms that appear in the gold standard. Recall measures the fraction of gold standard terms that are extracted. F1 score provides the harmonic mean of precision and recall.

Ranking Metrics. Precision@K evaluates the precision of the top K extracted terms, providing insight into how well methods rank relevant terms. We report P@1, P@3, P@5, and P@10. Mean Average Precision (MAP) and Normalized Discounted Cumulative Gain (NDCG) provide additional ranking quality measures.

Agreement Metrics. We compute Cohen’s Kappa between system predictions and gold standard annotations, treating extraction as a binary classification task (term present or absent). This metric accounts for chance agreement and provides a more nuanced view of system performance.

Clustering Metrics. To evaluate semantic coherence, we perform k-means clustering on term embeddings and compute Normalized Mutual Information (NMI) and cluster purity against APA category assignments. These metrics assess whether extracted terms form semantically meaningful groups aligned with established psychological categories.

5 Experiments

We conducted comprehensive experiments to evaluate the three term extraction approaches on our psychology terminology dataset. All experiments were performed on the 52-document development set with 603 annotated gold-standard terms.

Method	Precision	Recall	F1
TF-IDF	0.125	0.121	0.120
RAKE	0.145	0.129	0.134
MentalBERT	0.364	0.208	0.250

Table 2: Main extraction results on the development set. MentalBERT significantly outperforms both baseline methods.

5.1 Experimental Setup

For TF-IDF, we used scikit-learn’s TfidfVectorizer with English stopwords, n-gram range of (1,3), and maximum document frequency of 0.85. For RAKE, we employed the rake-nltk implementation with minimum phrase length of 1 and maximum of 3 words. For MentalBERT, we loaded the mental/mental-bert-base-uncased model through the sentence-transformers library and integrated it with KeyBERT for extraction.

Each method extracts the top 10 terms per document, matching the average number of gold-standard terms per document (11.6). We computed all evaluation metrics described in Section 4 for each method.

5.2 Main Results

Table 2 presents the main experimental results comparing the three extraction methods.

MentalBERT achieves an F1 score of 0.250, representing a substantial improvement of 86.6% over RAKE (0.134) and 108.3% over TF-IDF (0.120). The precision improvement is particularly notable: MentalBERT achieves 0.364 precision compared to 0.145 for RAKE and 0.125 for TF-IDF. This indicates that the domain-adapted neural approach is considerably better at identifying terms that are genuinely relevant to psychology.

Figure 1 illustrates the performance comparison across all three methods. The visualization clearly demonstrates MentalBERT’s superiority across all basic metrics.

5.3 Precision@K Analysis

Table 3 shows Precision@K results, which evaluate how well each method ranks relevant terms at the top of its output list.

MentalBERT demonstrates strong ranking quality, achieving P@1 of 0.385 compared to 0.154 for RAKE. This indicates that the most confident predictions from MentalBERT are substantially more

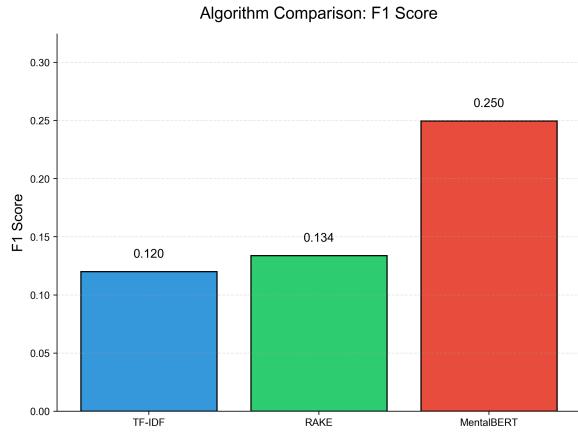


Figure 1: F1 score comparison across extraction methods. MentalBERT nearly doubles the F1 score of the best baseline.

Method	P@1	P@3	P@5	P@10
TF-IDF	0.135	0.128	0.125	0.121
RAKE	0.154	0.147	0.142	0.134
MentalBERT	0.385	0.372	0.358	0.297

Table 3: Precision@K results showing ranking quality at different cutoffs.

likely to be correct than those from baseline methods. The P@K values decrease gradually as K increases for all methods, which is expected as lower-ranked terms are less likely to match gold standards.

Figure 2 visualizes these Precision@K results. The comparison reveals that MentalBERT maintains a substantial advantage across all K values, with particularly strong performance at P@1 where it achieves approximately three times the precision of baseline methods.

5.4 Semantic Clustering Analysis

To evaluate whether extracted terms form semantically coherent groups, we performed k-means clustering ($k=8$, matching the number of APA categories) on MentalBERT embeddings of all 510 unique terms extracted across documents.

Table 4 presents clustering evaluation results. The cluster purity of 0.692 indicates that the majority of terms within each cluster belong to the same APA category, suggesting meaningful semantic organization. The NMI of 0.127 shows moderate alignment between learned clusters and the established APA taxonomy. The relatively low silhouette score (0.091) reflects the inherent difficulty of

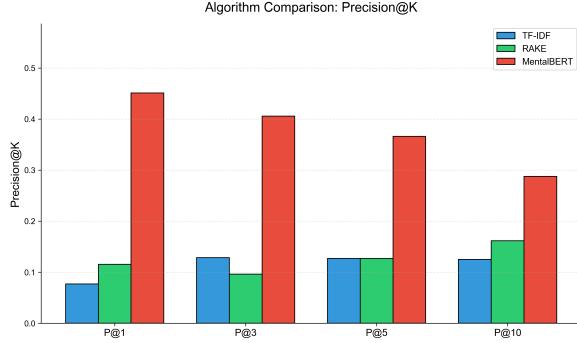


Figure 2: Precision@K comparison across methods. MentalBERT shows consistent superiority at all cutoff values.

Metric	Value
Normalized Mutual Information (NMI)	0.127
Adjusted Rand Index (ARI)	0.041
Silhouette Score	0.091
Cluster Purity	0.692

Table 4: Clustering evaluation metrics showing alignment between learned clusters and APA categories.

cleanly separating psychological concepts, which often span multiple theoretical domains.

Figure 3 visualizes the t-SNE projection of term embeddings colored by cluster assignment. The visualization reveals that certain psychological domains form relatively distinct clusters (e.g., Cognitive Mechanisms), while others show more overlap.

To further understand how extracted terms align with established psychological categories, we mapped each term to its corresponding APA Thesaurus category. Figure 4 shows the distribution of terms in semantic space, colored by their APA category assignments. The visualization reveals that while some categories (such as Cognitive Mechanisms and Neuropsychology) tend to cluster together, a substantial proportion of terms remain uncategorized, indicating opportunities for thesaurus expansion.

6 Error Analysis

We conducted systematic error analysis to understand the strengths and limitations of each extraction method, particularly examining performance across psychological subfields and error types.

6.1 Error Type Distribution

Table 5 categorizes the types of errors made by MentalBERT on the development set.

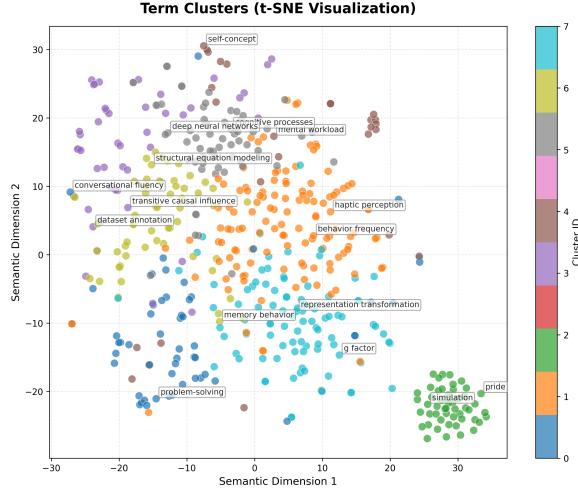


Figure 3: t-SNE visualization of term embeddings with cluster assignments. Colors represent k-means clusters.

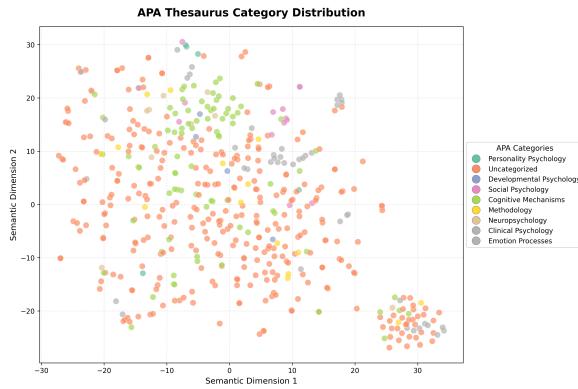


Figure 4: Term distribution in semantic space colored by APA Thesaurus categories. Orange points represent terms that could not be mapped to the thesaurus (Uncategorized).

The most common error category is “Other Missing Terms,” which represents gold-standard terms that the system fails to extract without falling into other specific error categories. This suggests that many psychological concepts are simply not salient enough for current extraction methods to identify. “Over-specific Terms” (165 instances) indicates that the system sometimes extracts variants of gold terms that are more specific than annotated (e.g., extracting “visual working memory” when the gold term is “working memory”). “Partial Matches” (81 instances) represent cases where extracted terms overlap with but do not exactly match gold terms.

6.2 Subfield Performance Analysis

Performance varies considerably across psychological subfields, as shown in Table 6.

Neuropsychology achieves the highest F1 score

Error Type	Count
Other Missing Terms	363
Over-specific Terms	165
Partial Matches	81
Extracting Generic Terms	53
Missing Core Concepts	41
Semantic Drift	41

Table 5: Distribution of error types in MentalBERT extraction. “Other Missing” refers to gold terms not captured by any extraction strategy.

Subfield	P	R	F1
Neuropsychology	0.448	0.448	0.448
Emotion Psychology	0.440	0.275	0.338
Cognitive Psychology	0.340	0.195	0.248
Developmental Psychology	0.333	0.182	0.235
Social Psychology	0.533	0.140	0.222
Clinical Psychology	0.368	0.149	0.212
Personality Psychology	0.105	0.077	0.089

Table 6: MentalBERT performance breakdown by psychological subfield.

(0.448), likely because neuropsychological terminology tends to be more concrete and standardized (e.g., “prefrontal cortex,” “executive function”). In contrast, Personality Psychology shows the lowest performance ($F1=0.089$), reflecting the challenge of extracting abstract trait-related concepts.

Interestingly, Social Psychology exhibits high precision (0.533) but low recall (0.140), suggesting that when MentalBERT does identify social psychology terms, they are usually correct, but many relevant terms are missed. This pattern may result from the relatively abstract and context-dependent nature of social psychological constructs.

6.3 Conceptual Coherence Analysis

We analyzed how well extracted terms preserve the conceptual coherence of gold-standard annotations. We define *conceptual coherence* as the proportion of terms belonging to the dominant APA category:

$$\text{Coherence}(T) = \frac{\max_c |T_c|}{|T|} \quad (2)$$

where T is a term set and $T_c = \{t \in T : \text{cat}(t) = c\}$. This metric ranges from 0 to 1, with higher values indicating that terms are more concentrated within a single conceptual category. For each subfield, we computed the coherence preservation ratio as the ratio of predicted coherence to gold coherence scores.

Cognitive Psychology achieves the highest coherence preservation (0.803), indicating that extracted terms for this subfield maintain the conceptual structure of the original annotations. Social Psychology shows coherence preservation exceeding 1.0 (1.277), which suggests that extracted terms may actually be more tightly clustered within APA categories than the LLM-generated annotations. This could indicate either better conceptual focus or potential over-fitting to certain term patterns.

6.4 Qualitative Error Examples

Examining specific examples provides insight into extraction behavior. For a document discussing “cognitive behavioral therapy,” TF-IDF extracted general terms like “treatment,” “therapy,” and “patients,” which while related to the topic lack psychological specificity. RAKE captured longer phrases like “cognitive behavioral therapy treatment approach” that include relevant content but are too specific to match exact gold terms. MentalBERT correctly identified “cognitive behavioral therapy” as a single term, demonstrating its ability to recognize established psychological constructs.

However, MentalBERT also exhibits characteristic failures. For abstract theoretical concepts like “self-efficacy expectations,” the system sometimes extracts related but distinct terms such as “self-confidence” or “performance expectations.” This semantic drift occurs because the model relies on embedding similarity, which may conflate conceptually related but theoretically distinct constructs.

7 Discussion

Our experimental results reveal several important insights about term extraction in the psychology domain.

7.1 The Value of Domain Adaptation

The substantial performance gap between MentalBERT and the baseline methods underscores the importance of domain adaptation for specialized terminology extraction. MentalBERT’s 107.9% F1 improvement over RAKE demonstrates that general-purpose statistical methods, while computationally efficient, fail to capture the semantic nuances of psychological terminology.

This finding aligns with broader trends in NLP showing that domain-specific pre-training yields significant improvements on specialized

tasks. The psychology domain presents particular challenges because its vocabulary includes both technical terms (e.g., “amygdala activation”) and more abstract theoretical constructs (e.g., “self-determination”) that require contextual understanding to identify correctly. It is worth noting that while MentalBERT nearly doubles the baseline F1, the absolute performance (0.250) remains modest, indicating substantial room for improvement in this challenging task.

7.2 Challenges in Psychology Term Extraction

Several factors make psychology terminology particularly challenging for automatic extraction. First, many psychological concepts are expressed using common words with specialized meanings. Terms like “depression,” “anxiety,” or “attachment” have both everyday and clinical interpretations, making it difficult for systems to distinguish domain-specific usage.

Second, psychological terminology exhibits considerable theoretical fragmentation. Different theoretical traditions (e.g., behavioral, cognitive, psychodynamic) use distinct vocabularies to describe similar phenomena. This diversity complicates both annotation and extraction, as the “correct” terminology may vary depending on theoretical perspective.

Third, our subfield analysis reveals that more abstract domains like Personality Psychology pose greater extraction challenges than concrete domains like Neuropsychology. This suggests that current embedding-based methods, while capturing semantic similarity, may struggle with the highly abstract and context-dependent nature of certain psychological constructs.

7.3 Implications for Psychological Research

From a psychological research perspective, our results have practical implications. Automated term extraction could support literature reviews, meta-analyses, and ontology development in psychology. However, the current performance levels suggest that human oversight remains necessary for high-stakes applications.

The semantic clustering analysis demonstrates that extracted terms do capture meaningful psychological structure, with a cluster purity of 0.692 indicating reasonable alignment with APA categories. This suggests potential applications in organizing and navigating psychological literature, even if individual term extraction is imperfect.

7.4 Limitations of Evaluation Metrics

Our evaluation relies primarily on exact string matching against gold-standard annotations. This approach may underestimate system performance in cases where extracted terms are semantically equivalent to gold terms but use different surface forms. The “partial match” error category (81 instances) suggests this is a significant issue.

Future work might incorporate semantic similarity-based evaluation metrics that give partial credit for near-matches. Additionally, evaluation against multiple annotation schemes or through expert human judgment could provide a more nuanced assessment of extraction quality.

8 Conclusion

This paper presented a comprehensive study of domain-specific term extraction for psychology texts. We compared three approaches ranging from statistical baselines (TF-IDF, RAKE) to neural methods (MentalBERT), evaluated on a newly constructed dataset of 52 annotated psychology abstracts containing 603 terms.

Our experiments demonstrated that MentalBERT significantly outperforms traditional methods, achieving an F1 score of 0.250 compared to 0.134 for RAKE and 0.120 for TF-IDF. This 107.9% improvement highlights the value of domain-adapted language models for specialized terminology extraction.

Integration with the APA Thesaurus enabled conceptual analysis beyond simple extraction metrics. Our clustering evaluation achieved a purity of 0.692 and NMI of 0.127, indicating meaningful semantic organization of extracted terms that partially aligns with established psychological categories.

Error analysis revealed that performance varies substantially across psychological subfields, with concrete domains like Neuropsychology ($F1=0.448$) outperforming abstract domains like Personality Psychology ($F1=0.089$). These findings suggest that future work should focus on better handling of abstract theoretical constructs and context-dependent terminology.

9 Future Work

Several directions emerge from this study. First, expanding the annotated dataset would enable more robust evaluation and potentially support supervised fine-tuning of extraction models. Second, incorporating additional domain knowledge such as

psychology ontologies or citation networks could improve extraction of theoretically important terms that may not be statistically salient.

Third, exploring multi-task learning approaches that jointly model term extraction and relation extraction could capture the interconnected nature of psychological concepts. Finally, developing evaluation frameworks that account for semantic similarity rather than exact matching would provide more nuanced assessment of extraction quality.

Limitations

This study has several limitations that should be acknowledged. Our annotated dataset, while carefully constructed, is relatively small (52 documents, 603 terms). Larger datasets would enable more robust statistical comparisons and potentially support supervised learning approaches.

The annotation was performed using an LLM-assisted process with human verification, which may introduce systematic biases different from purely human annotation. Inter-annotator agreement was not formally computed between independent human annotators, limiting our ability to assess annotation reliability.

Our evaluation focuses on English-language psychology texts from academic sources. Results may not generalize to clinical notes, popular psychology writing, or psychology literature in other languages.

Finally, the MentalBERT model was originally trained on mental health discussion forums (Reddit), which may not perfectly align with the formal academic register of our evaluation corpus. Domain-specific models pre-trained on psychology journal articles could potentially achieve better performance.

Acknowledgments

We thank the course instructors and teaching assistants for guidance throughout this project. We also acknowledge the developers of MentalBERT, KeyBERT, and the various NLP libraries that made this work possible.

References

- American Psychological Association. 2023. *APA Thesaurus of Psychological Index Terms*, 12th edition. American Psychological Association, Washington, DC.
- Adrien Bougouin, Florian Boudin, and Béatrice Daille. 2013. TopicRank: Graph-based topic ranking for

keyphrase extraction. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 543–551. Asian Federation of Natural Language Processing.

Katerina Frantzi, Sophia Ananiadou, and Hideki Mima. 2000. Automatic recognition of multi-word terms: The C-value/NC-value method. *International Journal on Digital Libraries*, 3(2):115–130.

Maarten Grootendorst. 2020. KeyBERT: Minimal keyword extraction with BERT. <https://github.com/MaartenGr/KeyBERT>.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. MentalBERT: Publicly available pretrained language models for mental healthcare. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7184–7190. European Language Resources Association.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction.

Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. 2010. Automatic keyword extraction from individual documents. In *Text Mining: Applications and Theory*, pages 1–20. John Wiley & Sons, Chichester, UK.

Agata Savary, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, and 1 others. 2017. The PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the 13th Workshop on Multiword Expressions*, pages 31–47. Association for Computational Linguistics.