Electronic Supplementary Material: Appendix

1 Methodological Guidelines of Abstract Annotation

Examples of both classes are depicted in Table 1, while a formal description of the classification procedure is demonstrated in Equation 1.

$$\text{abstract}_j = \begin{cases} \text{PRO} & \text{if for all result}_i \in \text{abstract}_j : \text{result}_i = \text{positive} \\ \text{MNR} & \text{if result}_i \text{ exists} \in \text{abstract}_j : \text{result}_i = \text{negative} \\ \text{exclusion} & \text{otherwise} \end{cases} \tag{1}$$

Table 1
Example abstract results

Number	Type of Results	Abstract
Example 1	Mixed or Negative Results	'Despite a largely successful stress induction, results do not support a reliable influence of experimentally induced social stress – with or without subsequent performance feedback – on pain in women. Further, we found no clear association of pain modulation and changes in neuroendocrine or subjective stress responses. ' (Schneider et al., 2022)
Example 2	Mixed or Negative Results	'Analyses revealed that individuals with PTSD symptoms showed AB for trauma- and depression-related words; however, mode of administration did not significantly influence reaction times. ' (Wittekind et al., 2017)
Example 3	Positive Results Only	'Our findings demonstrated the interrelation between higher levels of TEI and lower levels of BID among girls and boys. Positive associations were found between better HRQoL, better intrapersonal and stress management abilities (subscales of TEI) and lower BID, as reflected by parental and self-reports.' (Pollatos et al., 2020)

Note. Positive elements are highlighted in green, negative elements in red.

Examples 1 and 2 in Table 1 are both classified as MNR. Although Example 1 mentions a 'largely successful stress induction', we categorize it as MNR. This is because, following Van den Akker et al. (2023), we disregard manipulation checks and checks of statistical

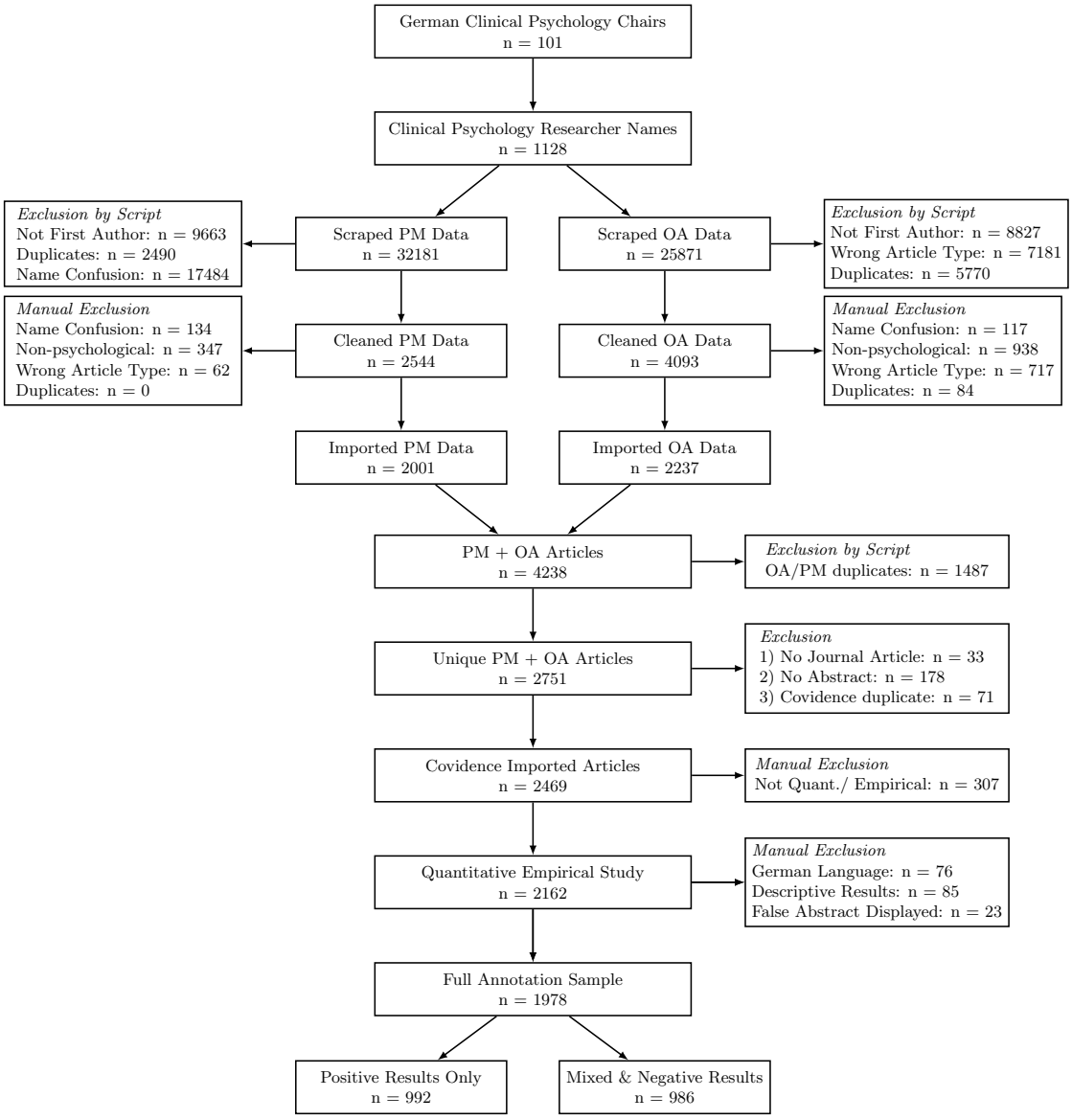
assumptions. As a result, the phrase ‘Despite a largely successful stress induction’ in the first section of this example is labeled as neither a positive result nor a negative result. Example 2 presents a positive result; however, since it also reports a negative result, it is categorized as MNR. Conversely, Example 3 is a PRO example as it exclusively reports positive results.

2 Data Collection for the Annotated Main Corpus

The data acquisition procedure for the annotated main corpus `MAIN` is shown in Figure 1.

Figure 1

Data acquisition procedure for the annotated `MAIN` corpus



Note. The `MAIN` corpus contains abstracts first-authored by German clinical psychology researchers and published between 2013 and 2022; OA = OpenAlex; PM = PubMed.

First, a list of all state university chairs ($n = 101$) which focused on clinical psychology, psychotherapy, or related fields in Germany, excluding laboratories in hospitals, was gathered. Chairs are located at 52 different universities from all 16 German federal states.

Second, all names, emails, and academic level (predoctoral, postdoctoral and professoral) of employed researchers were gathered from chair homepages, resulting in a list comprising 1,128 researcher names. 24 researchers were listed on more than one laboratory or chair web page. We extracted study meta data from (a) PubMed and (b) OpenAlex using R packages *rentrez* (Winter, 2017) and *openalexR* (Aria & Le, 2023), respectively. A *for loop* was used to iterate over the researchers names (n = 1,128) for both OpenAlex and PubMed. The exact search terms for OpenAlex and PubMed can be found in *Appendix* 4. This resulted in a total of 32,181 articles for PubMed and 19,662 articles for OpenAlex.

Third, the extracted OpenAlex and PubMed data were cleaned using a script. Initially, studies authored by researchers not from our predefined list were excluded (PM: n = 9,663; OpenAlex: n = 8,827), along with duplicate articles (PM: n = 2490; OpenAlex: n = 5,570). Notably, a substantial portion of the PubMed articles, 17,540 articles (55%), were attributed to being authored by just five researchers, consisting of two postdocs and three PhD candidates. This anomaly arose because PubMed has a built-in spelling correction function, which mistakenly included researchers with similar names in the corpus.

Upon investigating this discrepancy, we conducted a manual review of the total number of articles authored by these five researchers within our sample. The review revealed that only 56 articles were actually first-authored by these researchers, leading to the exclusion of the mistakenly included 17,484 entries from our dataset, which we categorized as 'name confusion'.

In the case of OpenAlex, additional cleaning steps were performed to exclude data points that did not represent journal articles according to OpenAlex (n = 6,209), or contained the terms 'systematic review','meta-analysis','comment', 'corrigendum', 'erratum' or 'correction' in the title section (OA: n = 972). This step was not necessary for PubMed as these article types could already be excluded through specifications in the search terms.

In the fourth step of our analysis, we manually screened both OpenAlex and PubMed study and journal titles using an Excel spreadsheet. We decided to exclude all apparent non-psychological studies from the corpus. We broadly defined these studies as those not reporting any association with the investigation of mental processes, emotions, or behaviors in humans. Following this criterion, we removed 347 articles from the PubMed dataset and 938 articles from the OpenAlex dataset that did not belong to our field of study. For the PubMed dataset, we identified and excluded 134 instances of name confusion (where researchers with similar or common names were incorrectly included) and 62 articles classified under wrong article types (e.g., review articles instead of original research). The exclusions for the OpenAlex dataset were more extensive. We found 117 instances of name confusion, 717 articles under incorrect article types, and additionally, identified and removed 84 duplicates.

In the fifth step, data from OpenAlex and PubMed were merged, with overlapping entries being excluded based on duplicate DOIs (n = 1,487). Standardized metadata were extracted using DOIs via the *rcrossref* library in R (Chamberlain et al., 2022). Crossref indicated that 33 of these entries did not represent journal articles, leading to their exclusion. Subsequently, abstracts were gathered using *Endnote* (Team, 2013). However, 178 entries lacked identifiable abstracts and were thus removed from the corpus.

The resulting dataset, comprising 2,540 entries, was imported into Covidence (Covidence, 2023). Covidence identified and removed an additional 71 duplicates based on available metadata. Following this step, the 'Title and Abstract Screening' sample in Covidence consisted of 2,469 entries.

After data preprocessing, two raters annotated a total of $n = 2469$ studies. During the title and abstract screening phase of our study, a total of $n = 307$ abstracts were excluded for not being quantitative or empirical in nature. Additional manual exclusions were as follows: $n = 76$ abstracts were removed because they were not written in English; $n = 85$ were excluded as they predominantly presented descriptive results without an indication of hypothesis-testing; $n = 23$ were dismissed from the data set due to issues with incorrect display of abstract texts, rendering them unsuitable for our study.

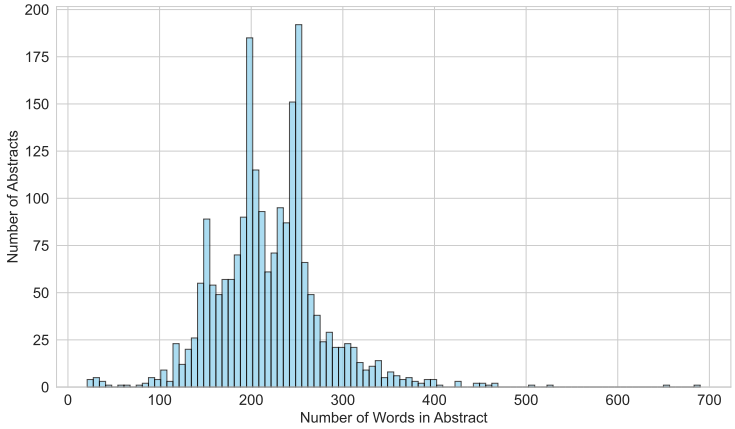
The resulting $n = 1978$ abstracts represent the development and in-domain data set for our classification task. This sample is referred to as `MAIN`.

3 Abstract Length

Figure 2 shows a histogram of the lengths of the abstracts in words utilized to fit our models. The mean abstract length was 221.1 (SD = 56.5), with a range from 31 to 690 words. In the figure, interestingly, three distinct local peaks are evident. These peaks correspond to the typical word length limits set for abstracts: 150, 200, or 250 words.

Figure 2

Histogram of abstract length (in words) for the `MAIN` *corpus*



Note. n = 1,978; number of bins = 100.

4 Search Terms

Search Terms In-Domain Data

OpenAlex: Scraping abstracts using the `openalexR` library:

```
library(openalexR)
oa_fetch(entity = "authors",
         display_name = name)


oa_fetch(
  entity = "works",
  author.id  = gsub(pattern = "https://openalex.org/",
                    replacement = "",
                    name),
  from_publication_date = "2013-01-01",
  to_publication_date = "2022-12-31",
  verbose = TRUE
)
```

PubMed: Scraping abstracts using the `rentrez` library:

```
library(rentrez)
entrez_search(
  db = "pubmed",
  api_key = api_key,
  tool = tool,
  email = email,
  term = paste0("(", name, '("NAME"[Author]) AND ("2013"[Date - Publication]:
  "2022"[Date - Publication]) NOT (Review[Publication Type]) NOT (Systematic
  Review[Publication Type]) NOT (Meta-Analysis[Publication Type]) NOT (Case
  Reports[Publication Type]) NOT (Editorial[Publication Type]) NOT
  (Letter[Publication Type]) NOT (Editorial[Publication Type]) NOT
  (News[Publication Type])')
)
```

Search Terms Out-of-Domain Data

Search terms for scraping abstracts from PubMed using the `rentrez` library:

- `VAL1`: 'Search: (Psychotherapy[MeSH Terms]) NOT (Germany[Affiliation]) Filters: Abstract, Randomized Controlled Trial'.

- `VAL2`: 'Search: (Psychotherapy[MeSH Terms]) AND ("1990"[Date - Publication] : "2012"[Date - Publication]) Filters: Abstract, Randomized Controlled Trial'

Search Terms & Preprocessing for Inference Data

Search terms for scraping abstracts from PubMed using the `rentrez` library:

- `INFER`: 'Search: (Psychotherapy[MeSH Terms]) AND ("1990"[Date - Publication]: "2022"[Date - Publication]) AND (Randomized Controlled Trial[Filter]) AND English[Language] NOT ((protocol[Title/Abstract]) OR (will[Title/Abstract]))'.

This search term deviates from the preregistered search term in two respects: First, we mistakenly preregistered the specification 'NOT ((protocol[Title/Abstract]) AND (will[Title/Abstract]))'; however, our intention was to exclude any study including the terms 'will' or 'protocol' to avoid having studies in the dataset that report future outcomes. Therefore, we replaced 'AND' with 'OR' to exclude all studies reporting either of these words in their title or abstract. Second, to ensure uniformity and comparability across the time span analyzed, we chose to restrict the analysis to the years with complete data, and thus only analyzed studies from 1990 to 2022, rather than extending the analysis to 2023. This decision was made since the analysis was conducted in the third quarter of 2023.

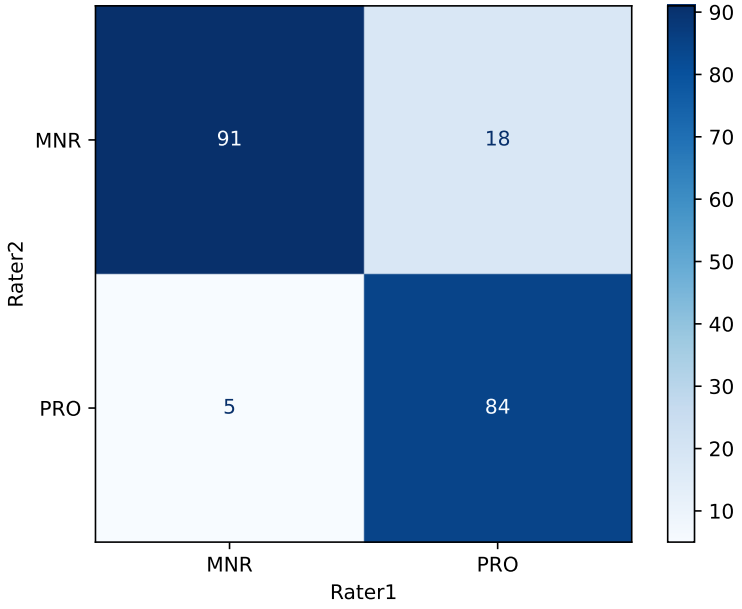
Using this search term, we were able to gather 20,862 abstracts from PubMed. For preprocessing we had to remove $n = 275$ observations without an abstract, $n = 93$ non-English abstracts identified with the R library *cld2* (Ooms, 2023), and $n = 61$ abstracts where the publication year was a missing value, and $n = 221$ where 2023 was the indexed publication year. Applying all preprocessing steps resulted in a total of 20,212 abstracts.

5 Interrater-Reliability

A confusion matrix demonstrating the disagreements between raters can be found in Figure 3. Disagreement was often associated with short adverbial constructs indicating exceptions such as 'but' or 'except'. For example, in some cases, two effects such as 'gender and remission' and 'SES and remission' were reported for Group *A*, *B* and *C*, but for Group *D*, only one effect, 'gender and remission', was reported

Figure 3

Confusion matrix for n = 198 labels between both raters

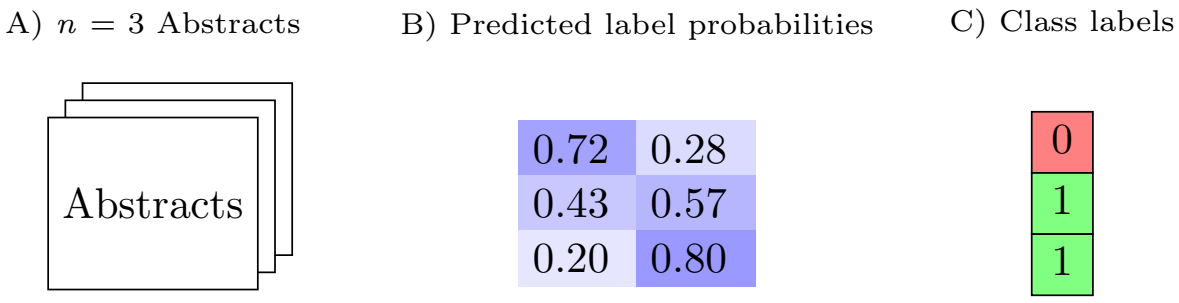


6 Supervised Learning

Supervised learning is an important subfield of machine learning, in which the primary objective is to construct a model from labeled data, enabling the prediction of unseen data (Mohri et al., 2018; Raschka et al., 2022). Based on a data set of labeled examples, supervised learning aims to 'learn' a function f that takes an input X to approximate the true class label y (Mohri et al., 2018). This data set, the so-called *training data*, is assumed to be labeled by a *supervisor* who knows the correct label for each instance. Thus, supervised learning is often referred to as 'label learning' (Raschka et al., 2022, p. 3). Using abstracts as an example, we can learn a function f that approximates the true class labels of y, 'MNR' or 'PRO', based on a corpus of labeled abstracts X. This function enables the prediction of whether new abstracts (referred to as *test data*) fall into either of the two categories. As illustrated in Figure 4, the learning process involves using $n = 3$ abstracts, which are processed by the supervised learning algorithm. The algorithm calculates the probability that each abstract belongs to class MNR or PRO. Lastly, the matrix of probabilities is compared with the labels. The 'learning' process essentially involves adjusting the parameters of the classifier to minimize the error between the predicted and observed labels.

Figure 4

Simple example of supervised learning based on abstracts



Note. The left column in the probability matrix corresponds to the probability that abstract$_j$ is of class MNR, while the right column in the matrix refers to the probability that abstract$_j$ is of class PRO. For the true class labels, '0' (red) corresponds to a true MNR label, and '1' (green) corresponds to a true PRO label.

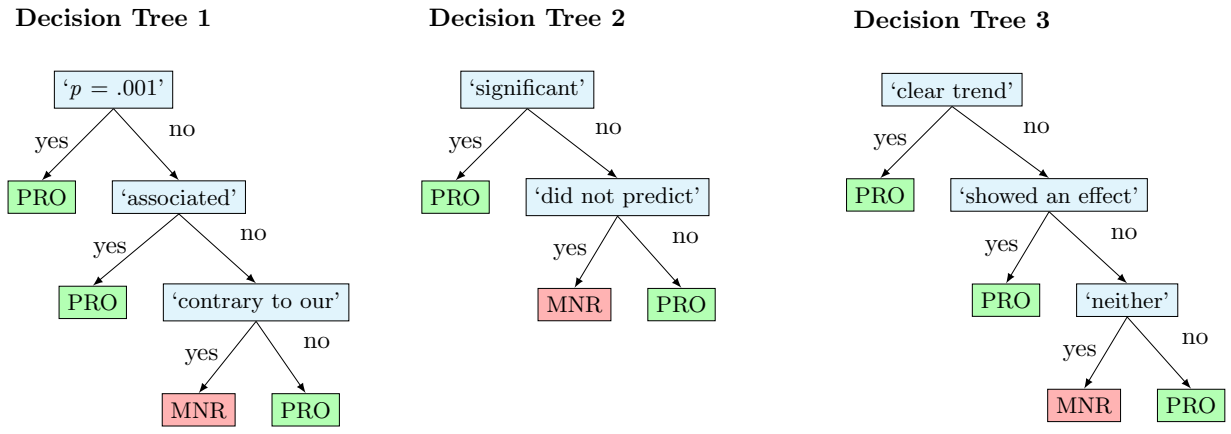
7 Random Forest

Random forests are a learning technique for classification and regression, which consist of a large number of *decision trees* (Breiman, 2001). Figure 5 shows a fictitious random forest of three example decision trees, which classify abstracts into the categories PRO or MNR based on *n*-grams such as 'clear trend' or 'did not predict'. Utilizing the features within the training data, the decision tree model learns a series of *n*-gram decisions to predict the class labels of the abstracts correctly. The decisions are strictly hierarchical. For instance, in *Decision Tree 2*, the class PRO is initially assigned if the term 'significant' appears within an abstract. This decision remains unchanged even if the phrase 'did not predict' is present. However, if 'did not predict' occurs but 'significant' is absent in the earlier step, the class MNR is assigned.

In random forest, an often-utilized decision scheme is the *majority vote*. Here, each individual decision tree in the ensemble 'votes' for a class, and the class with the most votes becomes the final predicted class for a given input (Fawagreh et al., 2014; Lam & Suen, 1997). Two primary advantages of random forest classifiers for our task include their adept handling of high-dimensional, noisy, sparse matrices in text classification (Islam et al., 2019)[1], and their robustness to overfitting (Breiman, 2001). Random grid parameters for the random forest

Figure 5

Example random forest of n *= 3 decision trees*



models used in our study are depicted in Table 2 and feature importances of the best model based on the training data are shown in 6.

[1] In text classification tasks, raw texts are converted into count-vectorized matrices. Here, columns correspond to specific *n*-grams, and rows correspond to text units. Given that specific words often appear in only a few units, these count-vectorized matrices can be considered as sparse matrices.

Table 2

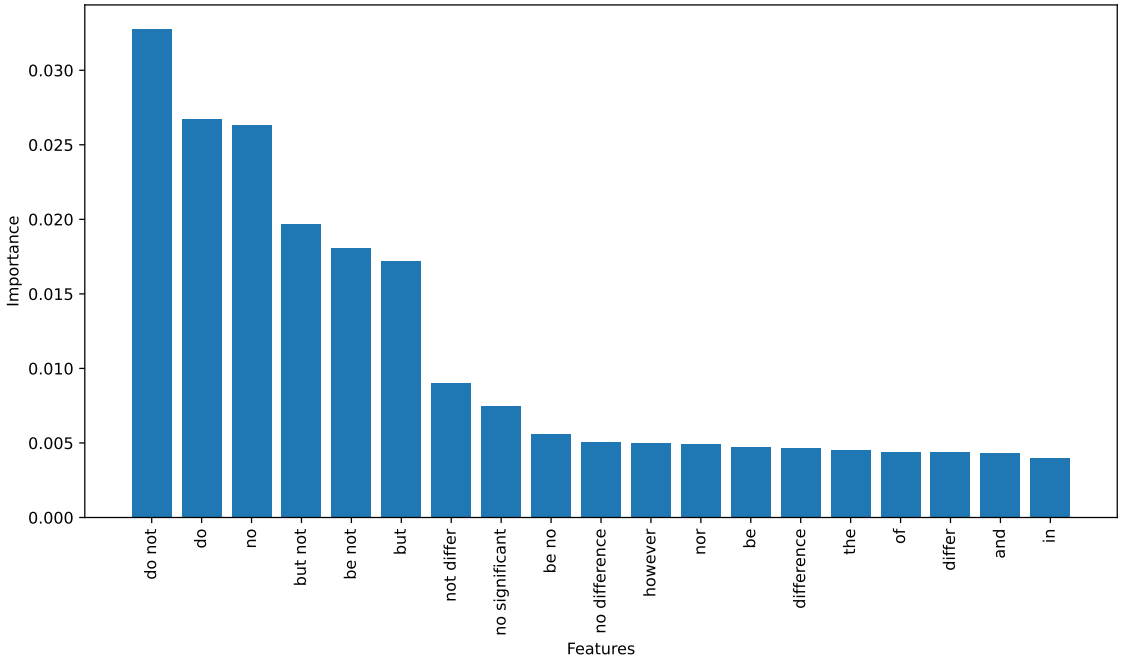
Random grid parameters for random forest models

Function	Parameter	Values	Best Value
CountVectorizer	max_features	[500, 1500, 2000]	2000
CountVectorizer	stop_words	[None, stopwords_adjusted]	None
CountVectorizer	ngram_range	[(1,2), (1,3)]	(1, 2)
RandomForestClassifier	n_estimators	[1000, 2000, 3000]	1000
RandomForestClassifier	max_features	[auto, sqrt]	auto
RandomForestClassifier	max_depth	[3, 10, 17, 24, 32, None]	32
RandomForestClassifier	min_samples_split	[2, 5]	2
RandomForestClassifier	min_samples_leaf	[1, 2, 4]	1
RandomForestClassifier	bootstrap	[True, False]	True

Note. Further informations on hyperparameters for `RandomForestClassifier` and `CountVectorizer` can be found in the documentation of *sklearn* (Pedregosa et al., 2011)

Figure 6

Feature importances for n *= 1780 abstracts (training set) from the best-performing random forest model*



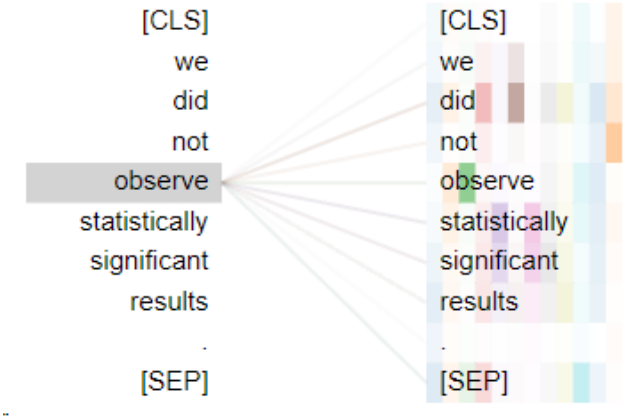
Note. Features were extracted using *sklearn*'s `.get_feature_names_out()` method (Pedregosa et al., 2011).

8 SciBERT

The strength of the connection between different parts of a text are captured by the so-called attention weights denoted as $\alpha_{i,j}$ (Raschka et al., 2022). These weights establish connections from inputs to outputs (contexts), necessitating two subscripts for the attention weights: j denotes the position of the input while i denotes the position of the output. Figure 7 illustrates the attention weights of $\alpha_{i,statistical}$ in the sentence 'We did not observe statistically significant results' within a single layer of our fine-tuned SciBERT model, utilizing *BertViz* (Vig, 2019).

Figure 7

Visualization of attention weights for the fine-tuned SciBERT model using BertViz



Note. Each line represents the attention from one input token j (on the left) to another output token i (on the right). The thickness of the line signifies the attention value $\alpha_{i,j}$, while the color differentiates between attention heads. The attention mechanism iteratively executes its computations, with each iteration referred to as an 'attention head' (Vig, 2019).

9 Validation Results

Table 3

Different metric scores for model evaluation of `MAIN`, `VAL1` and `VAL2` data sets

		Mixed & Negative Results			Positive Results Only		
	Accuracy	F1	Recall	Precision	F1	Recall	Precision
`MAIN`							
SciBERT	**0.864**	**0.867**	**0.907**	**0.830**	**0.860**	**0.822**	**0.902**
Random Forest	0.803	0.810	0.856	0.769	0.796	0.752	0.844
Extracted p-values	0.515	0.495	0.485	0.505	0.534	0.545	0.524
Extracted NL Indicators	0.530	0.497	0.474	0.523	0.559	0.584	0.536
Number of Words	0.475	0.441	0.423	0.461	0.505	0.525	0.486
`VAL1`							
SciBERT	**0.880**	**0.895**	**1.000**	**0.811**	**0.859**	**0.753**	**1.000**
Random Forest	0.787	0.810	0.883	0.747	0.758	0.685	0.847
Extracted p-values	0.547	0.547	0.532	0.562	0.547	0.562	0.532
Extracted NL Indicators	0.540	0.524	0.494	0.559	0.555	0.589	0.524
Number of Words	0.573	0.624	0.688	0.570	0.508	0.452	0.579
`VAL2`							
SciBERT	**0.847**	**0.857**	**0.945**	**0.784**	**0.835**	**0.753**	**0.935**
Random Forest	0.827	0.838	0.918	0.770	0.814	0.740	0.905
Extracted p-values	0.473	0.463	0.466	0.459	0.484	0.481	0.487
Extracted NL Indicators	0.527	0.510	0.507	0.514	0.542	0.545	0.538
Number of Words	0.547	0.414	0.329	0.558	0.630	0.753	0.542

Note. `MAIN` test data: $n = 198$ abstracts; `VAL1`: $n = 150$ abstracts; `VAL2`: $n = 150$ abstracts.

10 Inference Results

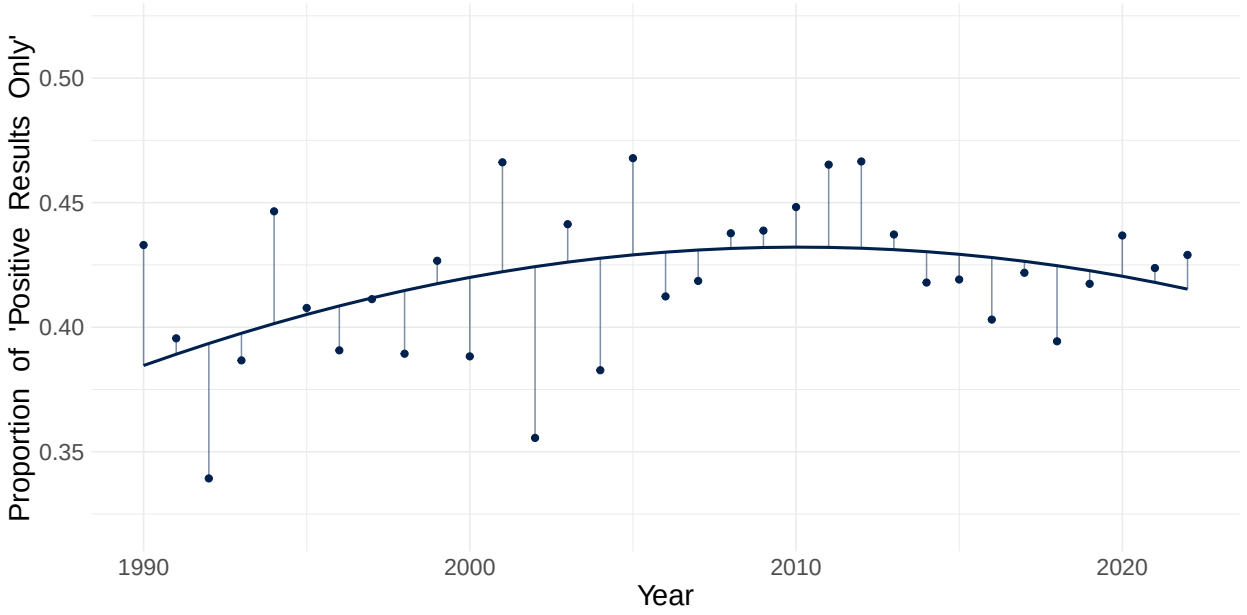
Table 4
Results from logistic regression models

Model	Outcome	Time Period	Term	b	SE	z	p
`MODEL1`	PRO	1990-2005	Intercept	-19.75	14.83	-1.33	.183
			Year (lin.)	9.70×10^{-3}	0.01	1.31	.191
`MODEL2`	PRO	2005-2022	Intercept	13.74	6.61	2.08	.038
			Year (lin.)	-6.96×10^{-3}	0	-2.12	.034
`MODEL3a`	PRO	1990-2022	Intercept	-3.17	3.74	-0.85	.396
			Year (lin.)	1.42×10^{-3}	0	0.77	.443
`MODEL3b`	PRO	1990-2022	Intercept	-1968.91	857.39	-2.3	.022
			Year (lin.)	1.96	0.85	2.29	.022
			Year (-quad.)	4.87×10^{-4}	0	2.29	.022
`MODEL3c`	PRO	1990-2022	Intercept	-28121.69	209690.78	-0.13	.893
			Year (lin.)	41.04	313.32	0.13	.896
			Year (-quad.)	2.00×10^{-2}	0.16	0.13	.898
			Year (cub.)	3.23×10^{-6}	0	0.12	.901
`MODEL3d`	PRO	1990-2022	Intercept	-1314.11	571.84	-2.3	.022
			Year (lin.)	0.98	0.43	2.3	.022
			Year (cub.)	-8.09×10^{-8}	0	-2.29	.022
`MODEL4a`	p<.05	1990-2022	Intercept	-13818.38	1467.5	-9.42	$< .001$
			Year (lin.)	13.74	1.46	9.41	$< .001$
			Year (-quad.)	3.42×10^{-3}	0	9.41	$< .001$
`MODEL4b`	p>.05	1990-2022	Intercept	-10667.57	3052.6	-3.49	$< .001$
			Year (lin.)	10.61	3.04	3.49	$< .001$
			Year (-quad.)	2.64×10^{-3}	0	3.49	$< .001$
`MODEL4c`	NLI(Pos.)	1990-2022	Intercept	-316.36	1139.74	-0.28	.781
			Year (lin.)	0.32	1.13	0.28	.781
			Year (-quad.)	7.91×10^{-5}	0	0.28	.78
`MODEL4d`	NLI(Neg.)	1990-2022	Intercept	2768.9	1337.68	2.07	.038
			Year (lin.)	-2.77	1.33	-2.08	.038
			Year (-quad.)	-6.92×10^{-4}	0	-2.09	.037

Note. b = unstandardized; SE = Standard Error; z = b/SE; PRO = Positive Results Only; lin. = linear; -quad. = negative quadratic; CI = Confidence Interval; NLI = Natural Language Indicator; Pos. = Positive Results; Neg. = Negative Results.

Figure 8

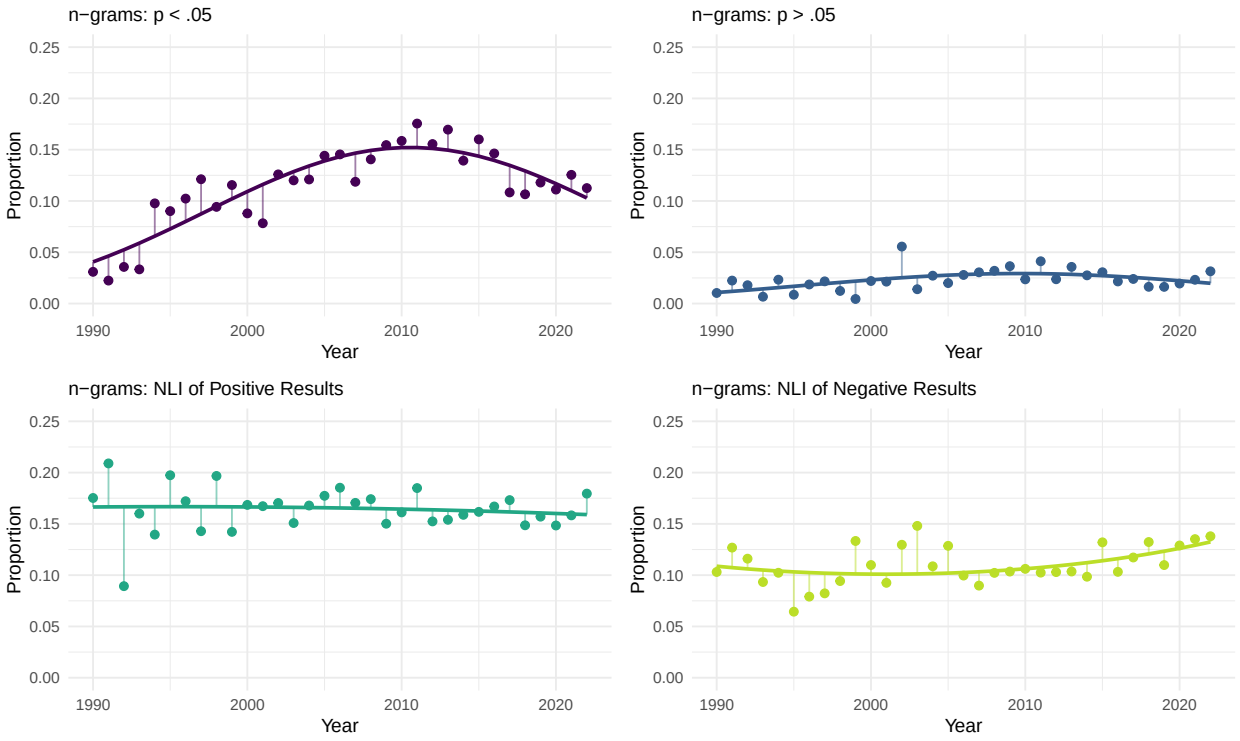
Predicted proportion of 'positive results only' in psychotherapy RCTs between 1990 and 2022 based on the SciBERT Model



Note. $n = 20{,}212$; Dots represent observed points. The bent line corresponds to the predicted proportion of PRO per year by `MODEL3b`.

Figure 9

Comparison of predicted proportions of positive and negative results in psychotherapy RCTs (1990-2022): rule-based approaches



Note. n = 20,212; NLI = Natural Language Indicator; dots represent observed values. Bent lines correspond to predicted proportions of PRO per year by $p < .05$ (`MODEL4a`), $p > .05$ (`MODEL4b`), natural language indicators of positive results (`MODEL4c`), and natural language indicators of negative results (`MODEL4d`).

Table 5

15 most frequent keywords per label for n *= 20,212 abstracts in* ***INFER***

	Mixed & Negative Results		Positive Results Only	
	Keywords	*n*	Keywords	*n*
1	Humans	8582	Humans	11623
2	Female	7293	Female	9997
3	Male	6500	Male	9038
4	Adult	5194	Adult	7240
5	Middle Aged	3881	Middle Aged	5515
6	Treatment Outcome	3496	Treatment Outcome	5158
7	Cognitive Behavioral Therapy	3063	Cognitive Behavioral Therapy	4268
8	Adolescent	2048	Adolescent	2886
9	Aged	1866	Aged	2601
10	Young Adult	1574	Young Adult	2136
11	Behavior Therapy	1397	Follow-Up Studies	2045
12	Follow-Up Studies	1341	Behavior Therapy	1893
13	Child	1121	Child	1403
14	Depression	983	Combined Modality Therapy	1284
15	Surveys and Questionnaires	903	Depression	1262

Note. Keywords for abstracts were gathered using R package *rentrez* (Winter, 2017).

11 SciBERT: Sensitivity Analysis for Alternative Breakpoints

We conducted a sensitivity analysis using piecewise linear regression on our inference data set to explore changes in the trend of positive results without splitting the data into two separate periods (1990-2005 and 2005-2022) and estimate whether an alternative breakpoint to 2005 is supported by the data. We utilized the `segmented` function from the *segmented* library (Fasola et al., 2018) to identify potential breakpoints in the time series. In our analysis, we specifically focused on identifying a single change point in our dataset, and thus set the total number of breakpoints to be estimated at `npsi` = 1. Based on the observation that only a small fraction, precisely 3.5%, of the data points were published before the year 1995, we defined the possible starting value for the breakpoint estimation at the year 1995 (`psi` = 1995). The model yielded the following findings:

- **Estimated Breakpoint:** The model demonstrated an estimated breakpoint in the year 2011 (Estimate = 2011, $SE = 2.55$).
- **Coefficients Before Breakpoint (year $\leq$ 2011):** The coefficient for Year_j before the breakpoint was positive ($b = 0.01$, $SE = 0.00$, $z = 2.34$, $p = .019$), indicating an increasing trend in positive results up to this point.
- **Change After Breakpoint (year $>$ 2011):** After the breakpoint, the coefficient turned negative (b = -0.02, $SE = 0.01$, z -3.07), suggesting a decrease in the trend of positive results post-2011.

These results of the break point analysis support the hypothesis of a shift in the trend of positive results, but place the breakpoint around the year 2011, rather than 2005.

12 SciBERT: Sub-trend Analysis

We expanded our analysis to investigate the predicted SciBERT trends more closely, focusing on three main aspects: regional influences, topic prevalence, and journal-specific tendencies. For each aspect, we fitted a logistic regression base model with Year_j and the negative square of Year_j ($-(\text{Year}^2)$) as predictors and additionally included further additive predictors. However, even after including these covariates for all three models, the effect of Year_j and ($-(\text{Year}^2)$) remained statistically significant and did not change the direction.

12.1 Regional Analysis

We first gathered country-specific information, enabling us to categorize studies based on four geographical regions: Europe, North America, Asia, and Rest of the World (Fanelli, 2012). We were able to gather information for regional affiliations for $n = 16,967$ first authors in our dataset. In summary as demonstrated by Figure 10 and Table 6, Asia showed the highest proportion of positive results, while research from Europe, North America, and the category 'Rest of the World' demonstrated a significant negative effect on the likelihood of being classified as PRO, compared to the reference region Asia.

Figure 10

Analysis of sub-trends in the likelihood of reporting positive results only as a function of world regions.

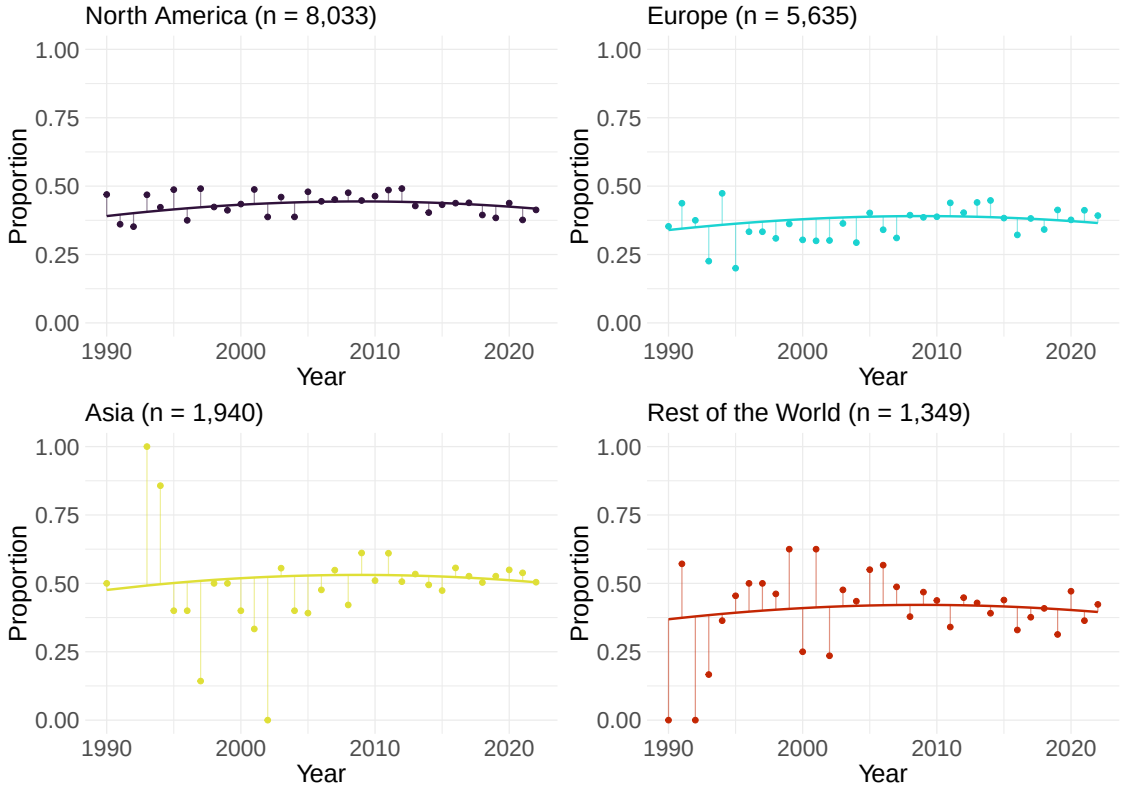


Table 6

Sub-trends in the likelihood of reporting positive results only as a function of world regions.

Term	*b*	*SE*	*z*	*p*
Intercept	-2507.06	958.65	-2.62	.009
Year (lin)	2.50	0.95	2.62	.009
Year (-quad.)	6.21×10^{-4}	0.00	2.62	.009
Europe	-0.57	0.05	-10.59	<.001
North America	-0.35	0.05	-6.71	<.001
Rest of the World	-0.44	0.07	-6.13	<.001

Note. The model's intercept, corresponds to year = 0, Year (-quad.) = 0 and region = 'Asia'.; *b* = unstandardized; *SE* = Standard Error; *z* = *b*/*SE*.

12.2 Topic-Specific Trends

We also explored the prevalence of specific topics within research abstracts. We used the occurrence of the four most-frequent psychology-specific terms ("cognitive", "depression", "anxiety", and "alcohol"; excluding stopwords and based on the frequency in our dataset) in our abstracts as binary predictors of PRO. In summary as demonstrated by Figure 11 and Table 7, the analysis of topic effects reveals that mentions of alcohol in abstracts significantly decrease the likelihood of being classified as having PRO. Similarly, cognitive-related terms show a significant negative impact on PRO classifications. However, the effects of depression and anxiety on the likelihood of positive results are not statistically significant.

Figure 11

Analysis of topic-specific sub-trends in the likelihood of reporting positive results only.

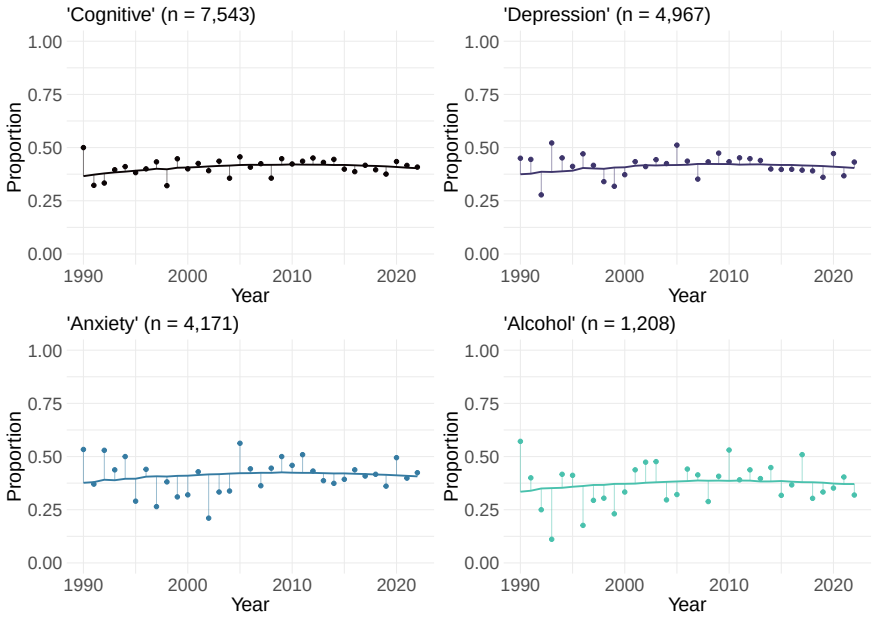


Table 7

Sub-trends in the likelihood of reporting positive results only as a function of reported topics.

Term	*b*	*SE*	*z*	*p*
Intercept	-2094.51	859.09	-2.44	.015
Year (lin)	2.08	0.86	2.44	.015
Year (-quad.)	5.18×10^{-4}	0.00	2.43	.015
Cognitive	-0.08	0.03	-2.50	.012
Depression	-0.04	0.04	-1.18	.237
Anxiety	-0.03	0.04	-0.85	.397
Alcohol	-0.22	0.06	-3.60	>.001

Note. b = unstandardized; *SE* = Standard Error; $z = b/SE$.

12.3 Journal-Specific Trends

Lastly, we delved into journal-specific trends by examining the 16 most frequent journals in our dataset.[2] By including the journal as a variable in our model, we aimed to discern any patterns that might suggest journal differences regarding positive/negative results.

Table 8

Journal-specific sub-trends in the likelihood of reporting positive results only.

Term	*b*	*SE*	*z*	*p*
Intercept	-4329.43	1812.27	-2.39	.017
Year (lin)	4.32	1.81	2.39	.017
Year (-quad.)	1.1×10^{-3}	0.00	2.39	.017
Addictive behaviors	0.11	0.22	0.50	.616
Amer. Acad. of Child & Adolescent Psychiatry	0.29	0.22	1.33	.185
Behavior therapy	0.25	0.18	1.37	.169
Behaviour research and therapy	-1.1×10^{-3}	0.22	-0.01	.996
Drug and alcohol dependence	-0.02	0.21	-0.08	.938
Journal of affective disorders	0.33	0.23	1.44	.150
Journal of anxiety disorders	0.32	0.17	1.84	.066
Journal of consulting & clinical psychology	0.41	0.22	1.89	.058
Journal of substance abuse treatment	0.13	0.22	0.61	.541
PloS one	0.19	0.22	0.85	.396
Psychological medicine	0.16	0.21	0.74	.462
Psychotherapy & psychosomatics	0.04	0.22	0.20	.844
The American journal of psychiatry	0.24	0.21	1.16	.248
The British journal of psychiatry	0.20	0.21	0.95	.340
The Journal of clinical psychiatry	0.01	0.23	0.05	.961

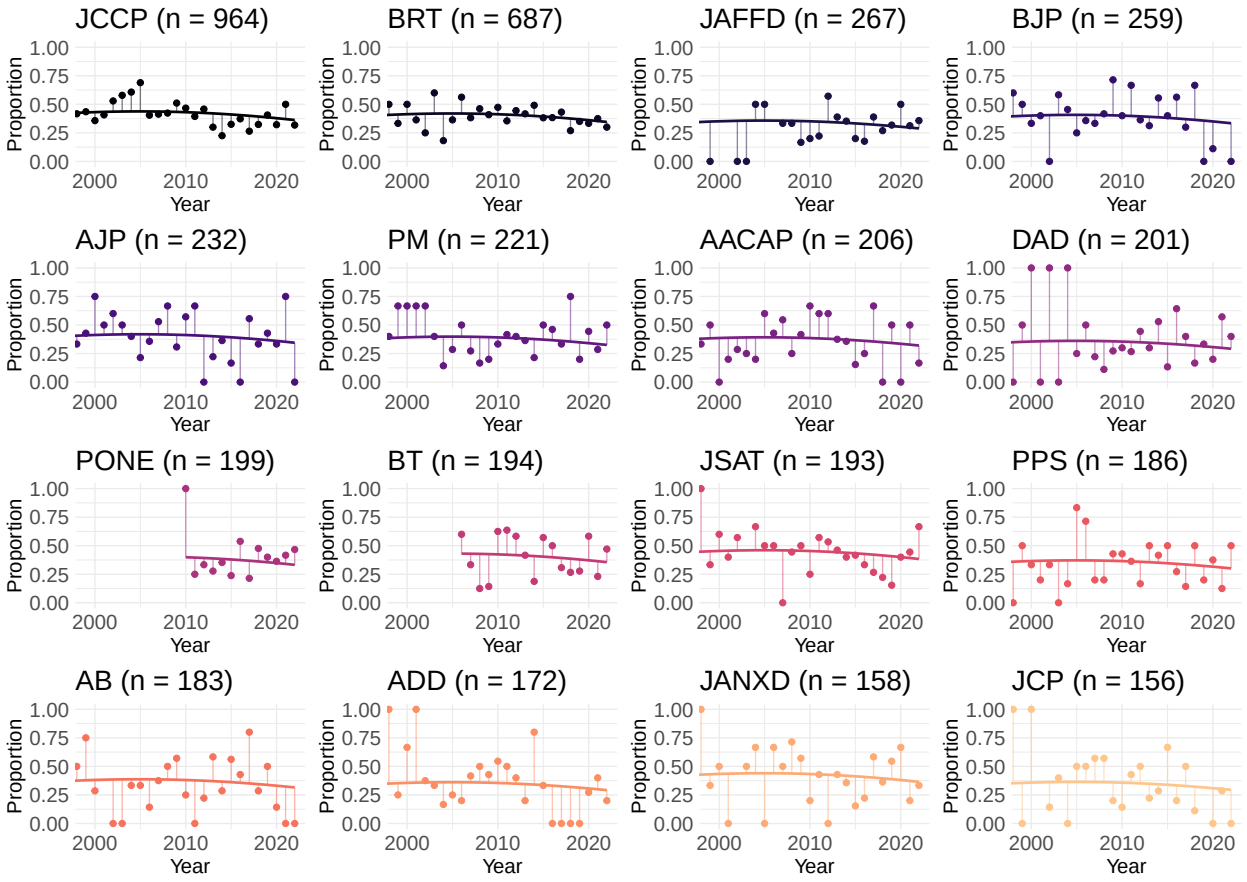
Note. The model's intercept, corresponds to year = 0, Year (-quad.) = 0 and journal = 'Addiction (Abingdon, England)'; b = unstandardized; *SE* = Standard Error; $z = b/SE$.

[2] We selected 16 journals because this number still allows for a clear and effective graphical representation (in a 4 x 4 grid).

Figure 11 and Table 7 demonstrate the result of this model. In summary, just like for the other models, the effect of Year_j and (Year^2)) were statistically significant, but no statistically significant effects were identified for any of the journals.

Figure 12

Analysis of journal-specific sub-trends in the likelihood of reporting positive results only.



Note. **Note.** Journal abbreviations are as follows: JCCP = Journal of Consulting & Clinical Psychology; BRT = Behaviour Research and Therapy; JAFFD = Journal of Affective Disorders; BJP = The British Journal of Psychiatry; AJP = The American Journal of Psychiatry; PM = Psychological Medicine; AACAP = American Academy of Child & Adolescent Psychiatry; DAD = Drug and Alcohol Dependence; PONE = PLoS One; BT = Behavior Therapy; JSAT = Journal of Substance Abuse Treatment; PPS = Psychotherapy & Psychosomatics; AB = Addictive Behaviors; ADD = Addiction; JANXD = Journal of Anxiety Disorders; JCP = Journal of Clinical Psychiatry.

13 SciBERT: Error Analysis

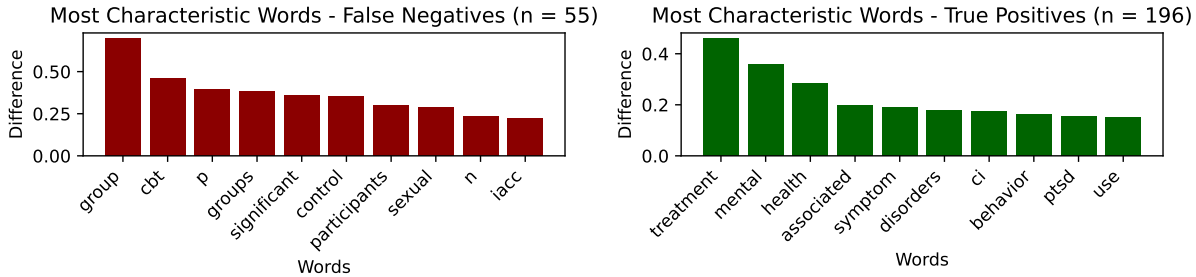
13.1 Misclassification and frequently reported words

In order to conduct an error analysis of the SciBERT model, we analyzed the most characteristic words for false negatives (FN), false positives (FP), true positives (TP), and true negatives (TN) from the test and validation sets (total of $n = 498$ studies) by calculating word frequencies within each category. We then adjusted these frequencies by the total word count to identify disproportionately common words in each classification outcome. Specifically, we compared false negatives with true positives as demonstrated by Figure 13 (manually annotated as class PRO), and false positives with true negatives as shown by Figure 14 (manually annotated as class MNR). Drawing definitive conclusions from this analysis is challenging because many of the words characteristic of each classification were often used not only in the results sections or conclusions but also in other parts of the abstracts.

However, an interesting pattern observed relates to the improvement of control groups, as indicated by the terms 'group', 'control', and 'groups'. When an abstract mentioned that experimental groups improved and control group scores remained the same, we did not categorize this as a negative result. Nonetheless, SciBERT appeared to process this as a negative result. For example, one of the false negatives reported that the score 'significantly improved in the experimental group, but not in the control group'. Another pattern characteristic of false negatives was the presence of 'p'. Consequently, we examined the p-values in false negatives and found that all p-values for the false negatives were below .05, which are typically considered positive results. Furthermore, 'significant' is also indicative of false-negatives. Similar to the terms 'group', 'control', and 'groups', the phrase 'no significant difference' was observed in several instances to describe change in the treatment but not in the control group. These instances were predicted as MNR but were annotated as PRO.

Figure 13

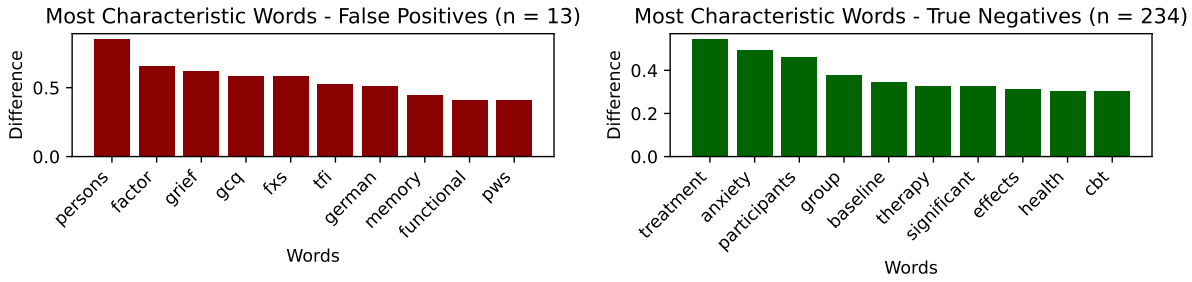
Comparison of the most characteristic words between false negatives and true positives.



The most characteristic words of the false positives do not seem to reveal certain words typically found in results or conclusion sections indicative of misclassification. Interestingly, 'CBT' (abbreviation for cognitive behavioral therapy) was identified as characteristic of both correct negatives and false negatives. This suggests that SciBERT tends to overpredict abstracts mentioning 'CBT' as belonging to the class MNR.

Figure 14

Comparison of the most characteristic words between false positives and true negatives.



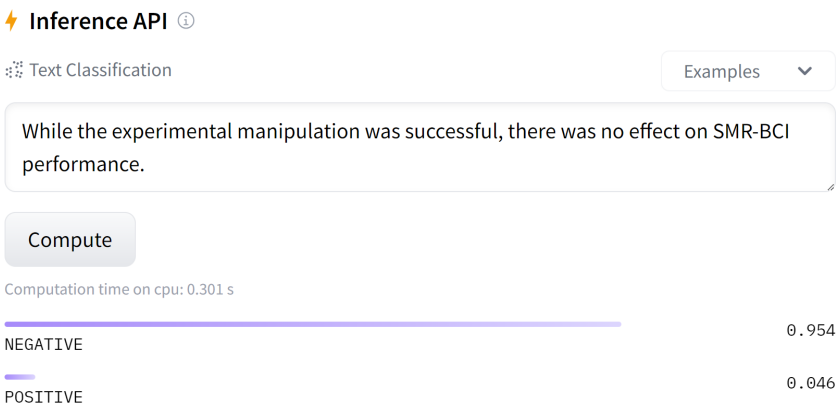
13.2 Manipulation checks

Following Van den Akker et al. (2023), we disregarded manipulation checks and checks of statistical assumptions, but we did not explicitly code whether a study mentioned statistical assumptions or manipulation checks. Nonetheless, we investigated whether our annotation strategy SciBERT was appropriately fine-tuned to down-weight the relevance of successful/unsuccessful manipulation checks. In the MAIN dataset, we identified 58 instances where the term "manipulation" was used. A brief manual analysis of the first 11 abstracts that mentioned "manipulation" revealed that 8 of these 11 abstracts included remarks about the success of the experimental manipulation in their results. Since SciBERT, unlike RF, is capable of processing linguistic context, we demonstrate using `Huggingface` that SciBERT accurately incorporated our annotation strategy and ignored experimental manipulation.

Example 1: Mixed and Negative Results

For example, as shown in Figure 15, in the statement, 'While the experimental manipulation was successful, there was no effect on SMR-BCI performance.' SciBERT correctly predicted the class as 'negative' (=MNR). The subphrase 'was no effect on SMR-BCI performance' was appropriately utilized in predicting this class, while the successfulness of the manipulation was ignored.

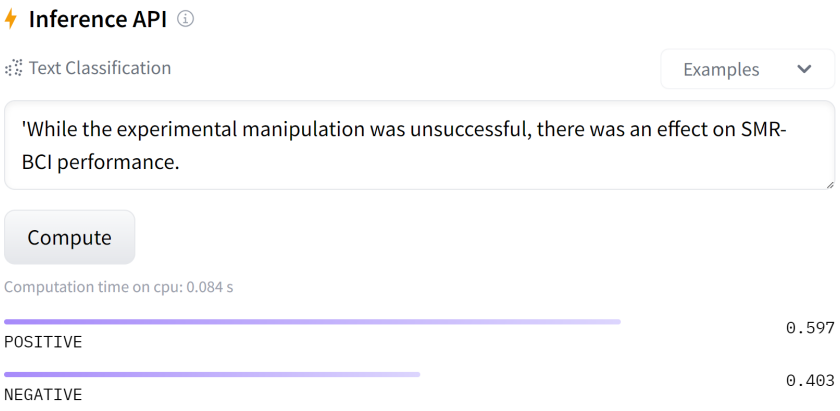
Figure 15
`HuggingFace`*: Successful manipulation check + negative result.*



Example 2: Positive Results Only

In the second example, as shown in Figure 16, 'While the experimental manipulation was unsuccessful, there was an effect on SMR-BCI performance.' the predicted class is 'positive' (=PRO). Also, here the subphrase 'was an effect on SMR-BCI performance' was higher weighted than the experimental manipulation. However, it is noteworthy that the MNR class also achieves a considerable prediction value (=0.403) in this case.

Figure 16
`HuggingFace`*: Unsuccessful manipulation check + positive result.*



We incorporated both examples to the *Inference API* on the model card of our `HuggingFace` repository.

References

Aria, M., & Le, T. (2023). *Openalexr: Getting bibliographic records from 'openalex' database using 'dsl' api*. https://github.com/ropensci/openalexR

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5–32.

Chamberlain, S., Zhu, H., Jahn, N., Boettiger, C., & Ram, K. (2022). Rcrossref: Client for various crossref apis [R package version 1.2.0]. https://CRAN.R-project.org/package=rcrossref

Covidence. (2023). Covidence review software. www.covidence.org.

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, *90*(3), 891–904. https://doi.org/10.1007/s11192-011-0494-7

Fasola, S., Muggeo, V. M., & Kuchenhoff, H. (2018). A heuristic, iterative algorithm for change-point detection in abrupt change models. *Computational Statistics*, *33*(2), 997–1015.

Fawagreh, K., Gaber, M. M., & Elyan, E. (2014). Random forests: From early developments to recent advancements. *Systems Science & Control Engineering: An Open Access Journal*, *2*(1), 602–609.

Islam, M. Z., Liu, J., Li, J., Liu, L., & Kang, W. (2019). A semantics aware random forest for text classification. *Proceedings of the 28th ACM international conference on information and knowledge management*, 1061–1070.

Lam, L., & Suen, S. (1997). Application of majority voting to pattern recognition: An analysis of its behavior and performance. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, *27*(5), 553–568.

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). *Foundations of machine learning*. MIT press.

Ooms, J. (2023). *Cld2: Google's compact language detector 2* [https://docs.ropensci.org/cld2/ (docs) https://github.com/ropensci/cld2 (devel) https://github.com/cld2owners/cld2 (upstream)].

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pollatos, O., Georgiou, E., Kobel, S., Schreiber, A., Dreyhaupt, J., & Steinacker, J. M. (2020). Trait-based emotional intelligence, body image dissatisfaction, and hrqol in children. *Frontiers in Psychiatry*, *10*, 973.

Raschka, S., Liu, Y. H., Mirjalili, V., & Dzhulgakov, D. (2022). *Machine learning with pytorch and scikit-learn: Develop machine learning and deep learning models with python*. Packt Publishing Ltd.

Schneider, S., Pauli, P., Lautenbacher, S., & Reicherts, P. (2022). Effects of psychosocial stress and performance feedback on pain processing and its correlation with subjective and neuroendocrine parameters. *Scandinavian Journal of Pain*.

Team, T. E. (2013). *Endnote* (Version EndNote X9) [64 bit]. Philadelphia, PA, Clarivate.

van den Akker, O. R., van Assen, M. A., Bakker, M., Elsherif, M., Wong, T. K., & Wicherts, J. M. (2023). Preregistration in practice: A comparison of preregistered and non-preregistered studies in psychology. https://osf.io/preprints/metaarxiv/fhdbs/

Vig, J. (2019). Bertviz: A tool for visualizing multihead self-attention in the bert model. *ICLR workshop: Debugging machine learning models*, *23*.

Winter, D. J. (2017). *Rentrez: An r package for the ncbi eutils api* (tech. rep.). PeerJ Preprints.

Wittekind, C. E., Muhtz, C., Moritz, S., & Jelinek, L. (2017). Performance in a blocked versus randomized emotional stroop task in an aged, early traumatized group with and without posttraumatic stress symptoms. *Journal of behavior therapy and experimental psychiatry*, *54*, 35–43.