Preprint: Natural Language Processing for Publication Bias Research - Using Machine Learning to Study Positive Results in Clinical Psychology Research Abstracts

Louis Schiekiera, Jonathan Diederichs, and Helen Niemeyer

Clinical Psychological Intervention, Freie Universität Berlin

Author Note

Correspondence concerning this preprint should be addressed to Louis Schiekiera, Clinical Psychological Intervention, Department of Education and Psychology, Freie Universität Berlin, Schloßstraße 1, 12163 Berlin, E-mail: l.schiekiera@fu-berlin.de

Abstract

Background: This study addresses the gap in machine learning tools for publication bias research by evaluating the performance of SciBERT and random forest in classifying results in clinical psychology abstracts. **Methods:** Over 1,900 abstracts were annotated into two categories: 'positive results only' and 'mixed or negative results'. Model performance was evaluated on three benchmarks. The best-performing model was utilized to analyze trends in over 20,000 psychotherapy study abstracts. **Results:** SciBERT outperformed all benchmarks and random forest in in-domain (accuracy: 0.86) and out-of-domain data (accuracy: 0.85-0.88). The trend analysis revealed non-significant effects of publication year on positive results for 1990-2005, but a significant decrease in positive results between 2005-2022. When examining the entire time-span, significant positive linear and negative quadratic effects were observed. **Discussion:** Machine learning could support future efforts to address publication bias by identifying negative results in large data sets. The fine-tuned SciBERT model was deployed for public use.

Keywords: publication bias, metascience, negative results, machine learning, natural language processing, text classification, SciBERT

1 Introduction

Systematic overrepresentation of positive results, commonly referred to as *publication bias*, affects the validity of research and leads to severely inflated effect sizes (Ferguson & Brannick, 2012). Apart from resulting in biased accumulation of findings, publication bias leads to both replication problems (Renkewitz & Heene, 2019) and waste of research resources (Mlinarić et al., 2017). Research on trends in publication bias has yielded mixed results, with studies using either *manual classification* or *rule-based algorithmic classification* for the estimation of the proportion of positive results in the literature (De Winter & Dodou, 2015; Fanelli, 2010; Leggett et al., 2013; Monsarrat & Vergnes, 2018; Pautasso, 2010).

Manual classification has been the standard in metascientific studies examining positive result proportions (Fanelli, 2010, 2012; Scheel et al., 2021). A notable contribution to the study of trends in publication bias was made by Fanelli (2012), who examined changes in statistical significance in over 4,600 papers based on the first reported hypothesis in a study and found an increase in positive results by over 22% between 1990 and 2007 across most scientific disciplines (Fanelli, 2012). However, drawbacks of manual classification are the extensive financial, time and intellectual resources needed for research synthesis (Marshall & Wallace, 2019).

More efficient, rule-based automated classification of study results emerged in the 2010s employing two strategies analyzing *n-grams* (De Winter & Dodou, 2015; Jager & Leek, 2014; Pautasso, 2010). N-grams can be described as linguistic units of sequences of n consecutive words or fragments in a text (Brown et al., 1990). The first strategy involves classification based on predefined n-grams of *natural language*. Typical natural language indicator (NLI) n-grams utilized in publication bias studies are e.g. 'no significant difference' for negative results, and 'significant difference' for positive results (Pautasso, 2010). The second strategy relies on classification using predefined n-grams of *statistical parameters* such as p-values like '$p >$ ' or '$p <$ ' (De Winter & Dodou, 2015; Jager & Leek, 2014) or the analysis of reported confidence intervals (Monsarrat & Vergnes, 2018).

Pautasso (2010) conducted a large-scale analysis of abstracts from 1970 to 2008 across multiple disciplines using a simple rule-based classification algorithm targeting NLIs of positive and negative results. Using only a few n-grams as markers, Pautasso (2010) observed that abstracts reporting significant differences grew more quickly than those indicating non-significant findings. Building on Pautasso (2010), De Winter and Dodou (2015) combined rule-based classification using n-grams of NLIs and statistical parameters and found that '$p < .05$' increased more slowly than '$p > .05$', whereas typical NLIs of positive results showed only a modest increase compared to NLI of negative results.

Despite their efficiency, predefined rules often capture only a limited array of expressions representing positive or negative results (Ioannidis, 2005). Moreover, the linguistic context in which these n-grams are presented is not processed by rule-based approaches.

When considering natural language, some abstracts might present statistically significant findings that are inconsistent with hypotheses. Additionally, when it comes to statistical parameters that depend solely on p-values, certain preliminary tests, such as Levene's test, lead to misclassifications, since here statistically significant results indicate assumption violations (Wells & Hintze, 2007) and not positive results. These problems led Ioannidis (2014, p.34) to state that: 'The abstracts of the crème de la crème of the biomedical literature are a mess. No fancy informatics script can sort out that mess. One still needs to read the papers'.

1.1 Letting Machines Learn from Annotations

Recent advancements in machine learning (ML), particularly in natural language processing (NLP), have enhanced the automation of research synthesis tasks such as study selection and data extraction (Marshall & Wallace, 2019). Supervised ML models trained on large sets of annotated data can efficiently process extensive textual datasets, addressing limitations of manual and rule-based methods (Beltagy et al., 2019). ML models, especially transformer-based architectures, can interpret the context of words, enabling the consideration of linguistic context (Devlin et al., 2018; Vaswani et al., 2017). However, they also face challenges, such as the potential for overfitting (Raschka et al., 2022), the demand for computational resources (Zimmer et al., 2023), and dependency on annotated datasets (Beltagy et al., 2019).

1.2 Positive Results in Clinical Psychology

High rates of positive results between 84–97% are identified in general for psychological studies (Fanelli, 2012; Open Science Collaboration, 2015; Scheel et al., 2021; Sterling, 1959), whereas for psychiatry and clinical psychology even up to 100% are observed (Rossignol & Frye, 2012).[1]

Publication bias in clinical psychological research can lead to misinformation among the public with respect to the efficacy of treatment options (Hopwood & Vazire, 2018). Moreover, productivity losses through mental health issues and treatment costs represent a substantial proportion of health-economic costs (Knapp & Wong, 2020) and biased data in the literature are used extensively in clinical decision analyses (Begg & Berlin, 1989).

1.3 Open Science

Over the past two decades, open science research practices, such as replication and registered reports, have been implemented increasingly in psychological research, and are both associated with higher rates of negative results (Open Science Collaboration, 2015; Scheel et al., 2021). The influence of open science practices on the proportion of positive results in clinical psychology is currently unknown. However, in a review, Tackett et al. (2019) suggested that open science practices are more pronounced in non-

[1] Notably, all studies (100% of 115) examined by Rossignol and Frye (2012) on the relationship between oxidative stress and autism indicated positive results.

clinical subdisciplines of psychology such as social and personality psychology compared to clinical psychology.

Following observations of Fanelli (2012), we assume that the proportion of positive results in clinical psychological research is linearly increasing between 1990 and 2005. Moreover, we assume that the publication of the seminal article, ‘Why Most Published Research Findings are False’ (Ioannidis, 2005), marked a shift in open science practices as it brought significant attention to the issue of false positives (Peterson & Panofsky, 2023). It captured attention in ways previous efforts had not and promoted the expansion of open science movements, and subsequent calls for methodological reforms (Open Science Collaboration, 2015; Peterson & Panofsky, 2023). Therefore, post-2005, we anticipate a decline in the proportion of positive results through to 2022 also in clinical psychology.

1.4 Present Study

This study addresses the lack of ML tools for analyzing publication bias and investigates shifts in positive result reporting in psychotherapy studies. To this aim, 1,978 abstracts authored by clinical psychology researchers affiliated with German universities and published in the past 10 years were categorized into two classes: ‘positive results only’ and ‘mixed or negative results’. We employed supervised ML models trained on human-annotated data from English-language abstracts. Specifically, we evaluated the performance of *Random Forest* and *SciBERT* and compared them with three benchmarks: classification based on NLIs, classification based on *p*-values and classification based on number of words. The models were out-of-domain validated using two sets of abstracts: (a) 150 abstracts of psychotherapy Randomized Controlled Trials (RCTs) written by researchers not affiliated with German universities and (b) 150 abstracts of psychotherapy RCTs from the period 1990-2012. Finally, the top-performing model was utilized to predict the prevalence of abstracts reporting ‘positive results only’ and ‘mixed or negative results’ for 20,212 unannotated abstracts from psychotherapy RCTs spanning the years 1990 to 2022.

2 Method

2.1 Abstract Annotation

In this study, the distinction between negative results and positive results is determined based on the presence/absence or the statistical significance/non-significance of a result (e.g. association, prediction, difference), rather than the direction (positive or negative) of the result. Like Van den Akker et al. (2023), we ignore manipulation checks and checks of statistical assumptions as well as descriptive results, when annotating abstracts. Furthermore, if a result is reported, but it is introduced by indicators of hypothesis-inconsistent results (e.g. ‘Contrary to our hypothesis’), it is also considered negative. However, since we want to train our model on units of abstracts, we have to consider that abstracts often contain multiple results. Therefore, we decided to annotate abstracts based on two categories: Positive Results Only (PRO) and Mixed or Negative Results (MNR).

Each result i in an abstract j's result section can be assigned to either class *positive* or class *negative* based on the assumptions made above. Given that abstracts often contain multiple results, we assign the class PRO to abstract$_j$ if all its results are of class *positive*. However, if abstract$_j$ contains at least one result i of class *negative*, it is labelled as being of class MNR. If neither of the conditions is met, we classify the abstract as an exclusion. Examples for both classes can be found in Appendix 1.

2.2 In-Domain Data

2.2.1 Data Collection

The sample of abstracts for building our models was derived from a subproject of our research group investigating negative results in publications of clinical psychology. Therefore, we gathered all *quantitative empirical original studies* first-authored by clinical psychology researchers affiliated with German universities from 2013 to 2022 in English language. Meta-analyses, reviews, editorials, comments, corrigendums, erratums, letters and qualitative studies were excluded. Abstracts were retrieved from PubMed and OpenAlex. While the other datasets consist only of abstracts from RCTs, we included both RCT and non-RCT abstracts in this dataset. The data acquisition procedure including several text mining and manual preprocessing steps can be found in Appendix 2.

The resulting $n = 1978$ abstracts represent the development and in-domain data set for our classification task. This sample is referred to as `MAIN`. 198 abstracts ($\tilde{1}0\%$ of 1978) were independently evaluated by both raters. Agreement in 88% of all abstracts and a $\kappa = .768$ suggests a reliable annotation process.[2]

2.3 Supervised Learning Pipelines

Code for training, evaluation, and prediction of SciBERT and the random forest pipelines is available on the project's *GitHub* repository (Schiekiera, 2023b). 81%, 9%, and 10% of the data were reserved for training, development and testing, respectively. A flowchart of the supervised learning pipelines is shown in Figure 1.[3]

2.3.1 Random Forest Pipeline

Random forests are a learning technique for classification and regression, which consist of a large number of *decision trees* (Breiman, 2001). For preprocessing in the random forest pipeline, we convert text to lowercase and apply lemmatization. In the subsequent step, the random forest pipeline transforms text data into a numerical format using

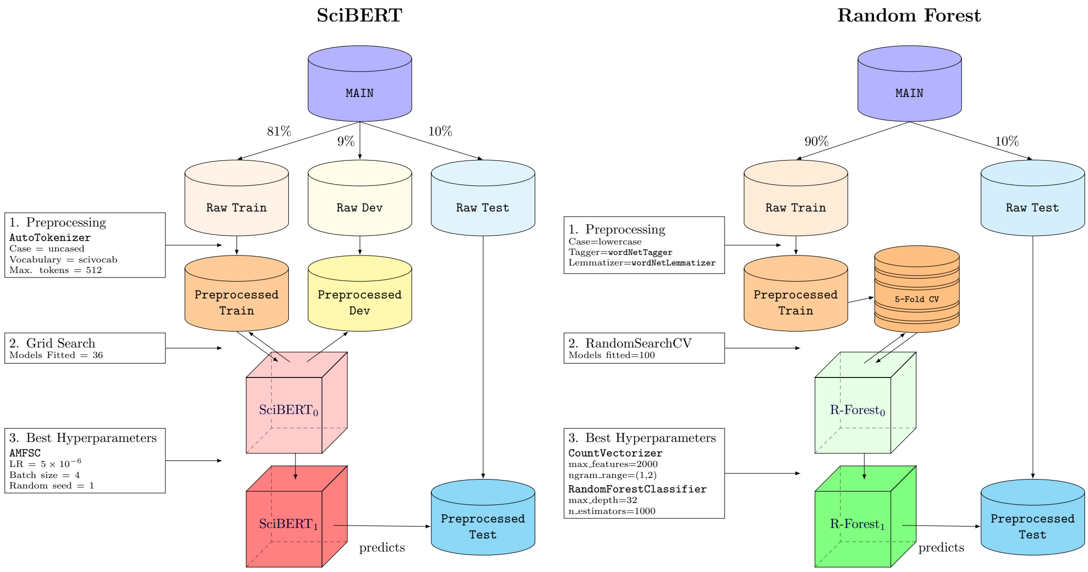
[2] Further information on interrater reliability can be found in Appendix 5.

[3] For further information on supervised learning see Appendix 6.

tokenization with `CountVectorizer`. It then employs a `RandomForestClassifier` for classification, both of which are implemented in scikit-learn (Pedregosa et al., 2011).[4].

Figure 1

Flowchart of the SciBERT and Random Forest pipelines



Note. LR = Learning Rate; AMFSC = AutoModelForSequenceClassification; CV = Cross Validation; Cylinders represent data sets and cubes represent models; SciBERT train data: n = 1,602; SciBERT dev data: n = 178; Random Forest train data: n = 1,780; Test data for both models: n = 198.

2.3.2 SciBERT Pipeline

SciBERT is a differently pretrained version of the BERT model (Devlin et al., 2018) and thus represents a Transformer (Vaswani et al., 2017). SciBERT's pretraining corpus consists of 18% papers from the computer science domain and 82% from the biomedical domain (Beltagy et al., 2019). In this pipeline we fine-tune SciBERT using our annotated abstracts. We first employ the `AutoTokenizer`, specifically in `allenai/ scibert_scivocab_uncased` settings to map words into numerical representations. SciBERT, like most BERT models, is limited to a maximum number of 512 input tokens (Beltagy et al., 2019). To optimize our model's performance, a comprehensive grid search was conducted using the `AutoModelForSequenceClassification` function from the *transformer* library. Hyperparameters under consideration are shown in Figure 1. The training was conducted with 3 epochs, and weight decay was set to $1e^{-2}$. The model showcasing the highest validation accuracy had a learning rate of $5e^{-6}$ and a batch size of 4.[5]

[4] For further information on random forest and the utilized hyperparameters in the random search see Appendix 7

[5] For further information on SciBERT and the self-attention mechanism see Appendix 8.

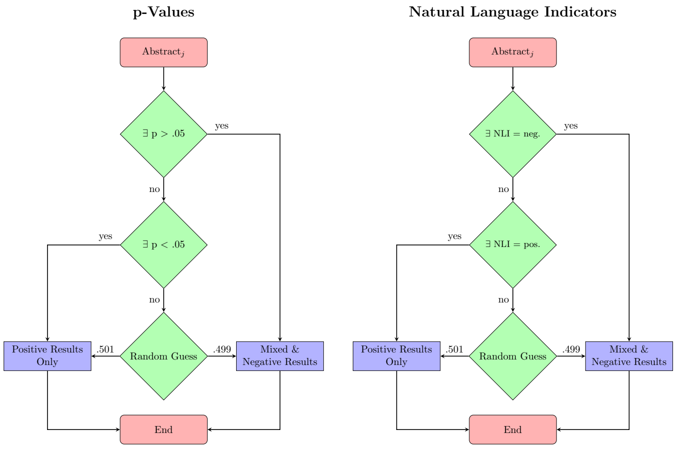
2.4 Benchmarks

2.4.1 Rule-Based Algorithms

The implementations of the rule-based approaches are based on extracted *p*-values (De Winter & Dodou, 2015) and on NLIs of negative and positive results (De Winter & Dodou, 2015; Pautasso, 2010) and are shown in Figure 2. The *n*-gram patterns are based on the search queries used by De Winter and Dodou (2015). We expanded the queries to also capture *p*-values between >.1 and .9. The full table of all queries can be accessed on the project's *GitHub* repository (Schiekiera, 2023b).

Figure 2

Algorithms for rule-based classification based on p*-values and natural language indicators of positive and negative results*



Note. NLI = Natural Language Indicators; ∃ = 'At least one .. '; .501 = proportion of 'positive results only' in training data; .499 = proportion of 'mixed or negative results' in the training data.

2.4.2 Naive Abstract Length Approach

Furthermore, we utilize a *naive abstract length approach* as an additional benchmark using a logistic regression classifier, which classifies the target based on abstract length in words.[6]

2.5 Out-of-Domain Data

For the out-of-domain data, we collected 300 abstracts from PubMed using the Medical Subject Headings (MeSH) term 'psychotherapy'. Since a random selection from journal abstracts might yield many studies not being empirical, we chose to focus only on RCTs to ensure we selected empirical studies. All sampled out-of-domain abstracts were checked to confirm that they contained quantitative results and were manually annotated as either MNR or PRO. To test for biases introduced by native German speakers in `MAIN`,

[6] Further descriptive information on abstract length can be found in Appendix 3.

we gathered 150 psychotherapy study abstracts, referred to as `VAL1`, which were first-authored by researchers *not* affiliated with a German university. To account for potential temporal biases, we sampled 150 psychotherapy study abstracts from 1990–2012. This sample is referred to as `VAL2`.[7]

2.6 Inference

2.6.1 Data

For our inference dataset, we aimed to predict the proportion of abstracts that reported only positive results from 1990 to 2022. The preregistration of our analysis can be found in this OSF Preregistration (for PubMed search terms and further details see Appendix 4). In total, we gathered 20,862 psychotherapy study RCT abstracts from PubMed, which resulted in a total of 20,212 abstracts after preprocessing.[8].

2.6.2 Modelling

We operationalized our outcome variable as the *result type* of an abstract. We coded abstracts with MNR as 0 and with PRO as 1. We then applied our best-performing model to predict the class labels (MNR, PRO) for all the abstracts obtained from this search using logistic regression in R. In the logistic regression models, we predict the probability that every result$_i$, within a psychotherapy study abstract$_j$ is positive (=PRO), using the publication year as a predictor. This relationship is defined by:

$$P(\text{abstract}_j = \text{PRO}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \times \text{Year}_j)}} \tag{1}$$

In this equation, $P(\text{abstract}_j = \text{PRO})$ is the probability that abstract j exclusively reports positive results. β_0 is the intercept, and β_1 represents the coefficient of the predictor, which in this case is Year$_j$. This equation represents the foundation for `M1` and `M2`, investigating the years 1990-2005 and 2005-2022, respectively.

Furthermore, `M3a` tests for a linear increase over the whole time-span by merging both data sets (1990-2005 & 2005-2022). `M3b` tests for an inverted U-shape effect by adding a second regression coefficient β_2, which is multiplied with the negative square of Year$_j$ ($-(\text{Year}^2)$). Introducing this negative quadratic term $-(\text{Year}^2)$ allows us to capture, in line with our hypothesis, trends of initial increases and subsequent decreases. Moreover, `M3c` introduces a cubic regression coefficient β_3 which is multiplied with the cubed term of Year$_j$ (Year^3). In `M4d` the quadratic term is removed and only the linear and the cubic term predict the outcome variable. Akaike Information Criterion (AIC) is used to compare `M3a`, `M3b`, `M3c` and `M3d`.

To compare longitudinal trends between predictions and rule-based approaches, it was also investigated whether the reporting of p-values $< .05$ (`M4a`), p-values $> .05$ (`M4b`),

[7] PubMed search terms for both data sets can be found in Appendix 4.

[8] For further information on preprocessing see Appendix 3

NLIs of positive results (`M4c`) and of negative results (`M4d`) varies as a function of Year_j and its negative quadratic term. For all models standardized betas (β) are reported. Reproducible code for all models is available on the project's *GitHub* repository (Schiekiera, 2023b).

3 Results

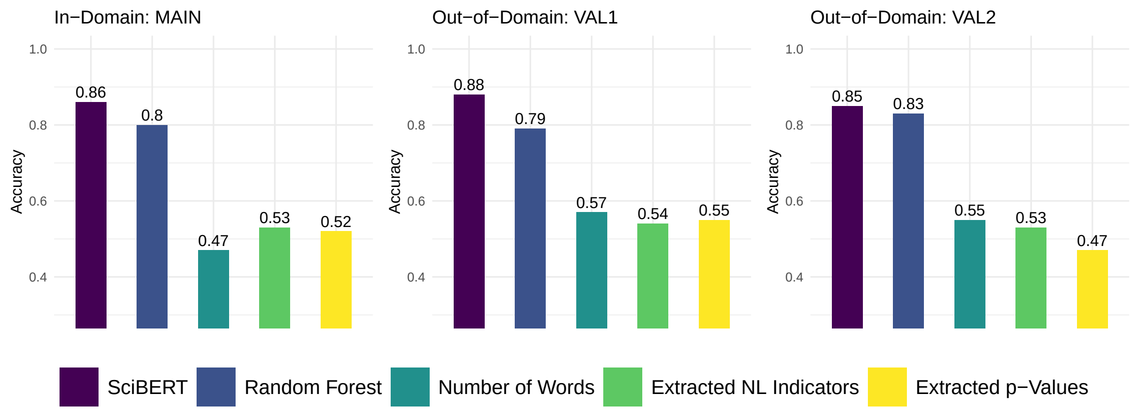
3.1 Validation

The labels of the $n = 1{,}978$ abstracts in the `MAIN` corpus were evenly divided between the PRO and MNR categories: 50% were annotated as PRO, while 50% were classified as MNR. Similarly, for `VAL1`, 51% were labeled as MNR and 49% as PRO, and for `VAL2`, 49% were annotated as MNR and 51% as PRO (both $n = 150$).

Accuracy scores of the classification models based on in-domain and out-of-domain data are illustrated in Figure 3. Further metrics can be found in Appendix 9. The SciBERT model outperformed the other models, achieving the highest accuracy of 0.86 for `MAIN` and similar scores for out-of-domain data (0.85-0.88). The random forest model, while not as proficient as the SciBERT, displayed solid performance with an accuracy of 0.80 for `MAIN` and robust accuracies for out-of-domain data (0.79-0.83). Rule-based classification based on extraction of p-values and predefined NLIs of positive and negative results, as well as the classification based on the number of words rendered results around the chance of random guessing for in-domain and out-of-domain data (between 0.47 and 0.57).

Figure 3

Comparing model performances across in-domain and out-of-domain data



Note. Colored bars represent different model types; Samples: `MAIN` test: $n = 198$ abstracts; `VAL1`: $n = 150$ abstracts; `VAL2`: $n = 150$ abstracts.

3.2 Deployment

The best-performing model, SciBERT, was deployed under the name 'NegativeResult-Detector' (Schiekiera, 2023a). It can be used via a graphical user interface for single abstract evaluations or for larger inference by downloading the model from *HuggingFace*, utilizing a script from the *GitHub* repository (Schiekiera, 2023b).

3.3 Inference

3.3.1 SciBERT Classifications

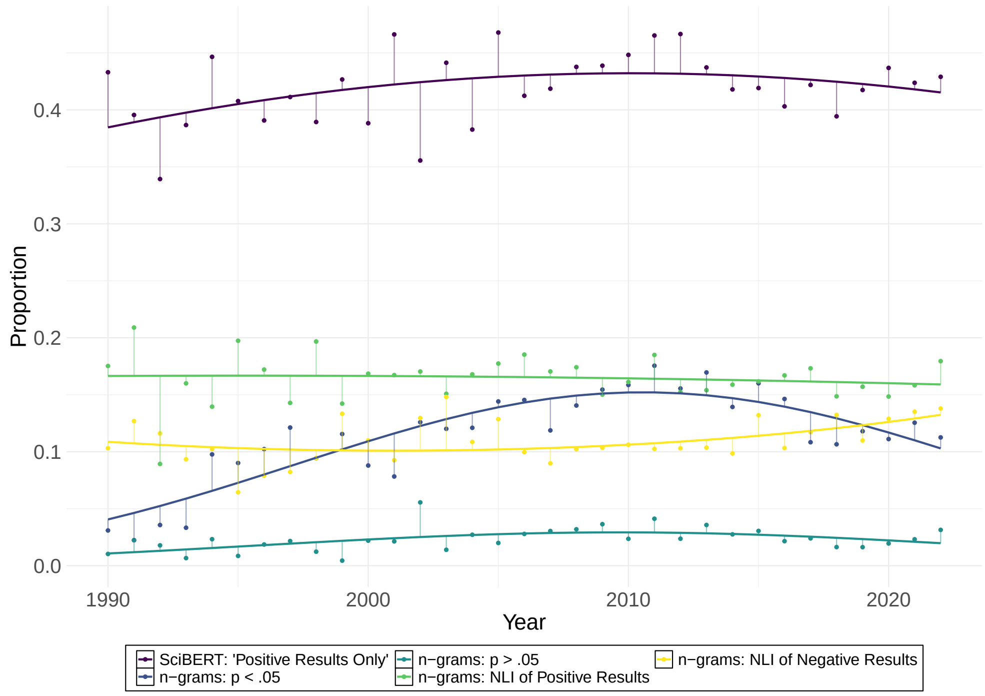
The evaluations of both in-domain and out-of-domain data indicate that SciBERT offers the best performance. Consequently, this model was employed for inference. For 1990-2005 we observed no statistically significant linear effect (`M1`: $\beta = 0.04$, $p = .191$). However, there was a negative statistically significant linear effect for publication year for the period 2005-2022 (`M2`: $\beta = -0.03$, $p = .034$). When merging both data sets no significant effect was found for publication year (`M3a`: $\beta = 0.01$, $p = .443$). When adding a negative quadratic term to the equation in `M3b`, we observed that both the effect of the linear term for year ($\beta = 15.02$, $p = .022$) and the effect for the negative quadratic term for year were significant ($\beta = 15.01$, $p = .022$). Introducing a further cubic term in `M3c`, yielded non-significant effects for all coefficients (linear: $\beta = 314.80$, $p = .896$; quadratic: $\beta = 614.81$, $p = .898$; cubic: $\beta = 300.02$, $p =.901$). However, `M3d` with only a linear and a cubic term showed significant effects for both regression terms (linear: $\beta = 7.52$, $p = .022$; cubic: $\beta = -7.51$, $p =.022$). When comparing all models spanning the years 1990-2022, `M3b` and `M3d` showed the best, but also identical AIC (`M3a`: AIC = 27563.61; `M3b`: AIC = 27560.34; `M3c`: AIC = 27562.33; `M3d`: AIC = 27560.34). Therefore, we furthermore employed the Bayes Information Criterion (BIC), which showed also identical indices for both models (`M3b`: BIC = 27584.08; `M3d`: BIC = 27584.08). For better interpretability we chose `M3b` for further consideration. Proportions of PRO per year and the predicted regression line of `M3b` in comparison the rule-based approaches are depicted in Figure 4. Despite strong fluctuations between 1990 and 2005, this model reflects the observed trend of PRO over time: an initial increase in PRO from a lower proportion in the early 1990s (1990 to 1992: $M = 0.39$, $SD = 0.05$) to a consistent relative peak in the early 2010s (2010 to 2013: $M = 0.45$, $SD = 0.01$). Following this peak, a modest decline led to a moderately high proportion of PRO in the early 2020s (2020 to 2022: $M = 0.43$, $SD = 0.01$). The lowest proportion of PRO was observed in 1992 (0.33) and the highest in 2005 (0.47).

3.3.2 Rule-Based Classifications

In line with the identified linear positive and negative quadratic trend of PRO, we found significant positive linear and negative quadratic effects for the presence of *n*-grams indicating '$p < .05$' (`M4a`. Linear: $\beta = 105.43$, $p < .001$; Quadratic: $\beta = 105.32$, $p < .001$) as well as for '$p > .05$' (`M4b`. Linear: $\beta = 81.41$, $p < .001$; Quadratic: $\beta = 81.37$, $p < .001$). However, no significant changes over time were identified for the presence of NLIs of positive results (`M4c`. Linear: $\beta = 2.42$, $p = .781$; Quadratic: $\beta = 2.44$, $p = .780$). NLI of negative results demonstrated significant negative effects for the linear

Figure 4

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



Note. n = 20,212; NLI = Natural Language Indicator; dots represent observed values. Bent lines correspond to predicted proportions of PRO per year by SciBERT (`M3b`), p < .05 (`M4a`), p > .05 (`M4b`), natural language indicators of positive results (`M4c`), and natural language indicators of negative results (`M4d`).

and the quadratic term (`M4d`. Linear: β = -21.25, p = .038; Quadratic: β = -21.33, p = .037). Thus, in contrast to the other models, the positive quadratic effect suggests a slight U-shaped rather than an inverted U-shape. Additional results can be found in Appendix 10.

4 Discussion

In summary, this study had three main objectives. First, we evaluated the reliability of our result classifier utilizing the annotated `MAIN` corpus. Second, we assessed the generalizability of our model by examining the performance of SciBERT on two additional annotated samples of psychotherapy RCT abstracts, which included both non-German samples and publications from earlier time periods (1990 - 2012). Third, we used SciBERT to predict the result type across an extensive collection of psychotherapy RCTs from 1990 to 2022.

4.1 Proportion of Mixed and Negative Results

Our study found proportions of MNR between 49-58% in the data. This contrasts with the lower negative result rates in psychology, reported as 4% to 34% in previous studies

(Fanelli, 2010; Scheel et al., 2021; Toth et al., 2021; van den Akker et al., 2023). This discrepancy can be explained by the fact that abstracts typically report several results. Therefore, the probability that at least one result$_i$ in abstract$_j$ is negative is higher than the probability for a single result$_i$ in abstract$_j$ to be negative.

4.2 Validation

The classification results underscore the potential of ML models in publication bias research. A central advantage over traditional rule-based methods (De Winter & Dodou, 2015; Pautasso, 2010) is the ability of ML to learn heterogeneous reporting styles of results. Only 9% of the abstracts in `MAIN` mentioned *p*-values, and only 14% utilized predefined NLIs of positive or negative results, despite all being quantitative. In 21% of abstracts, at least one rule-based *n*-gram was detected, leaving 79% where classifications would be left to random guessing. Yet, SciBERT and random forest stand out with their capacity to utilize extensive *n*-gram sets for predictions. They circumvent the limitations of depending on a narrow set of linguistic cues. While both the random forest and SciBERT models show solid performance, SciBERT's superiority in every metric underscores the advancements of NLP through the introduction of Transformer models using the self-attention mechanism to enhance processing of linguistic context.

4.3 Inference

SciBERT demonstrated superior performance in predicting both in-domain and out-of-domain data. Consequently, we utilized this model for our inference task, which aimed to detect patterns related to the prevalence of PRO in psychotherapy RCTs from 1990 to 2022. When examining the data linearly over the period 1990-2005, no significant effect for publication year was found for this period. However, a linear decrease in PRO was observed for 2005-2022. The absence of a linear increase in positive results during the 1990s and early 2000s deviates from the observation of Fanelli (2012) describing a substantial increase in positive findings from 1990 to 2007 across disciplines, including psychology and psychiatry. These differing outcomes may be attributable to methodological differences. Our study segmented abstracts into PRO and MNR, in contrast to Fanelli (2012), who focused on the statistical significance of the first reported hypothesis in full articles. Additionally, we specifically analyzed RCTs, while Fanelli (2012) analyzed all kinds of quantitative primary research. This difference between RCTs and other studies might stem from early awareness of publication bias in clinical trials (Dickersin et al., 1987). Furthermore, from the 1980s to the 2010s, psychotherapy RCTs in the US were particularly well funded in contrast to other research designs in psychotherapy research (Goldfried, 2016). Sufficient funding is often considered a protective factor against publication bias (Fanelli, 2012). However, our results could also indicate a trend difference in psychotherapy studies compared to other areas in psychology.

When combined, the data revealed significant quadratic and linear trends, depicting an increase in PRO during the 1990s, peaking in the early 2010s, and then declining. In line with our hypothesis, the highest proportion of positive results was observed in 2005 following the publication of Ioannidis (2005), but this value seemed rather an outlier than a consistent peak over time. However, the consistent peak in the early 2010s may be due to a 'time lag effect', indicating that research trends take time to manifest in publications as they slowly gain acceptance among researchers.

Furthermore, both '$p < .05$' and '$p > .05$' displayed significant positive linear and negative quadratic patterns over time, closely mirroring the PRO trends. Although De Winter and Dodou (2015) did not control for quadratic effects over time and analyzed trends across disciplines, we observed, similar to De Winter and Dodou (2015), an average increase of '$p < .05$'. Factors contributing to this increase might include a rise in questionable research practices and the growth of structured reporting including p reporting (De Winter & Dodou, 2015). Similarly, to the observations in the PRO category, the subsequent decline of '$p < .05$' after the early 2010s might reflect methodological discussions around open science in psychology. However, the increase of '$p > .05$' is likely to reflect a rise in structured abstracts as well. Surprisingly, the proportion of '$p > .05$' decreased despite the discourse around open science following 2010.

Moreover, NLIs of positive results did not show any statistically significant changes over time, which contrasts with De Winter and Dodou (2015) and Pautasso (2010). This discrepancy might arise because the 'p-value algorithm' can recognize the entire spectrum of p-values, but the set of NLIs is restricted to a narrow range, thus capturing only a fraction of the expressions indicating negative or positive results. However NLIs of negative results demonstrated a slight U-shaped increase over time, with particularly low proportions of NLIs of negative results in the mid 1990s. An increase in NLIs of negative results was also reported by De Winter and Dodou (2015) and Pautasso (2010).

4.4 Limitations

This study has three main limitations. First, abstracts instead of full texts were examined. This might result in missing out on details found in the full text, potentially leading to misclassifications. Additionally, it should be highlighted that the reporting standards may vary between abstracts and their corresponding full texts, as underscored by Assem et al. (2017).

Second, the choice to classify abstracts into two categories, PRO and MNR, might oversimplify the representation of abstract result sections. For instance, a study reporting several positive outcomes, but one negative outcome would still fall under the MNR class. A more nuanced approach could have entailed an ordinal classification, breaking down results into solely positive, mixed, and entirely negative. Alternatively, a metric method could have been adopted wherein the ratio of negative outcomes in an abstract is measured.

Third, we did not implement a strategy to differentiate between quantitative and non-quantitative studies, nor between descriptive and hypothesis-testing studies for `INFER`. To address this, our focus was set on RCTs, although it is worth noting that some RCTs rely on qualitative rather than quantitative methods (Nelson et al., 2015).

4.5 Conclusion

This study presented a novel approach in negative results detection using NLP. The robust performance of our models, especially SciBERT, demonstrates the potential for the use of ML in improving research synthesis tasks. Applying the SciBERT model to an extensive sample of psychotherapy RCTs, our study identified a trend of an initial increase in psychotherapy study abstracts reporting only positive results from the early 1990s to the early 2010s, which changed in the early 2010s to a subsequent decrease in the reporting of positive results only. Although the observed trends were small and strong fluctuations were evident in the early 1990s, this recent decrease in the reporting of positive results only represents a promising trend. This observation aligns with the growing awareness of open science topics in the clinical psychology research community in recent years. As demonstrated in this study, ML models are valuable tools for revealing such trends and could be crucial in future efforts to understand and address publication bias. Our methodological contributions and findings should encourage further investigations using ML models. Exploring more nuanced result type target variables beyond the binary classes presented in this study could provide deeper insight into this critical aspect of scientific research.

5 References

Assem, Y., Adie, S., Tang, J., & Harris, I. A. (2017). The over-representation of significant p values in abstracts compared to corresponding full texts: A systematic review of surgical randomized trials. *Contemporary Clinical Trials Communications*, *7*, 194–199.

Begg, C. B., & Berlin, J. A. (1989). Publication bias and dissemination of clinical research. *JNCI: Journal of the National Cancer Institute*, *81*(2), 107–115.

Beltagy, I., Lo, K., & Cohan, A. (2019). Scibert: Pretrained language model for scientific text. *EMNLP*.

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

Brown, P. F., Cocke, J., Della Pietra, S. A., Della Pietra, V. J., Jelinek, F., Lafferty, J., Mercer, R. L., & Roossin, P. S. (1990). A statistical approach to machine translation. *Computational linguistics*, *16*(2), 79–85.

De Winter, J. C., & Dodou, D. (2015). A surge of p-values between 0.041 and 0.049 in recent decades (but negative results are increasing rapidly too). *PeerJ*, *3*, e733.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *preprint arXiv: 1810.04805.*

Dickersin, K., Chan, S., Chalmersx, T., Sacks, H., & Smith Jr, H. (1987). Publication bias and clinical trials. *Controlled clinical trials*, *8*(4), 343–353.

Fanelli, D. (2010). "positive" results increase down the hierarchy of the sciences. *PloS one*, *5*(4), e10068.

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, *90*(3), 891–904.

Ferguson, C. J., & Brannick, M. T. (2012). Publication bias in psychological science: Prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. *Psychological methods*, *17*(1), 120.

Goldfried, M. R. (2016). On possible consequences of national institute of mental health funding for psychotherapy research and training. *Professional Psychology: Research and Practice*, *47*(1), 77.

Hopwood, C., & Vazire, S. (2018). Reproducibility in clinical psychology.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine*, *2*(8), e124.

Ioannidis, J. P. (2014). Discussion: Why "an estimate of the science-wise false discovery rate and application to the top medical literature" is false. *Biostatistics*, *15*(1), 28–36.

Jager, L. R., & Leek, J. T. (2014). An estimate of the science-wise false discovery rate and application to the top medical literature. *Biostatistics*, *15*(1), 1–12.

Knapp, M., & Wong, G. (2020). Economics and mental health: The current scenario. *World Psychiatry*, *19*(1), 3–14.

Leggett, N. C., Thomas, N. A., Loetscher, T., & Nicholls, M. E. (2013). The life of p:“just significant” results are on the rise.

Marshall, I. J., & Wallace, B. C. (2019). Toward systematic review automation: A practical guide to using machine learning tools in research synthesis. *Syst Rev*, *8*(1), 163.

Mlinarić, A., Horvat, M., & Supak Smolcić, V. (2017). Dealing with the positive publication bias: Why you should really publish your negative results. *Biochemia medica*, *27*(3), 447–452.

Monsarrat, P., & Vergnes, J.-N. (2018). The intriguing evolution of effect sizes in biomedical research over time: Smaller but more often statistically significant. *GigaScience*, *7*(1), gix121.

Nelson, G., Macnaughton, E., & Goering, P. (2015). What qualitative research can contribute to a randomized controlled trial of a complex community intervention. *Contemporary Clinical Trials*, *45*, 377–384.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Pautasso, M. (2010). Worsening file-drawer problem in the abstracts of natural, medical and social science databases. *Scientometrics*, *85*(1), 193–202.

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.

Peterson, D., & Panofsky, A. (2023). Metascience as a scientific social movement. *Minerva*, 1–28.

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.

Renkewitz, F., & Heene, M. (2019). The replication crisis and open science in psychology.

Rossignol, D. A., & Frye, R. E. (2012). A review of research trends in physiological abnormalities in autism spectrum disorders. *Molecular psychiatry*, *17*(4), 389–401.

Scheel, A. M., Schijen, M. R. M. J., & Lakens, D. (2021). An excess of positive results: Comparing the standard psychology literature with registered reports. *Advances in Methods and Practices in Psychological Science*, *4*(2).

Schiekiera, L. (2023a). NegativeResultDetector. https://doi.org/10.57967/hf/1146

Schiekiera, L. (2023b). *Repository: NegativeResultDetector*. https://github.com/schiekiera/NegativeResultDetector

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American Statistical Association*, *54*(285), 30–34.

Tackett, J. L., Brandes, C. M., King, K. M., & Markon, K. E. (2019). Psychology's replication crisis and clinical psychological science. *Annual review of clinical psychology*, *15*, 579–604.

Toth, A. A., Banks, G. C., Mellor, D., O'Boyle, E. H., Dickson, A., Davis, D. J., DeHaven, A., Bochantin, J., & Borns, J. (2021). Study preregistration: An evaluation of a method for transparent reporting. *Journal of Business and Psychology*, *36*, 553–571.

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.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Wells, C. S., & Hintze, J. M. (2007). Dealing with assumptions underlying statistical tests. *Psychology in the Schools*, *44*(5), 495–502.

Zimmer, M., Spiegel, C., & Pokutta, S. (2023). Sparse model soups: A recipe for improved pruning via model averaging. *arXiv preprint arXiv:2306.16788*.

6 Acknowledgments

Thanks to the Berlin University Alliance for funding. Special thanks to Max Zimmer and Jurek Bauer for their helpful comments regarding the implementation and validation of our machine learning models. Special thanks as well to Dr. Jochen Laubrock for supervising Louis Schiekiera's master thesis, which represents the foundation of this paper. Lastly, we appreciate the constructive feedback of Caspar Schauß, Sophie Lutter and Dr. Karoline Sachse regarding the methodology of our study.

7 Authorship

- Conceptualization: L. Schiekiera, H. Niemeyer
- Methodology: L. Schiekiera, H. Niemeyer
- Software: L. Schiekiera, J. Diederichs
- Validation: L. Schiekiera
- Formal Analysis: L. Schiekiera
- Investigation: L. Schiekiera, H. Niemeyer
- Resources: H. Niemeyer
- Data Curation: L. Schiekiera, J. Diederichs
- Writing – Original Draft: L. Schiekiera
- Writing – Review & Editing: L. Schiekiera, J. Diederichs, H. Niemeyer
- Visualization: L. Schiekiera
- Supervision: H. Niemeyer
- Project Administration: H. Niemeyer
- Funding Acquisition: H. Niemeyer

All authors approved the final version of the article.