Toward A Robust Method for Understanding the Replicability of Research

Ben Gelman,[1] Chae Clark,[1] Scott Friedman,[2] Ugur Kuter,[2] James Gentile[1]
[1]Two Six Labs, Arlington, VA, USA
[2]SIFT, Minneapolis, MN, USA
{ben.gelman, chae.clark, james.gentile}@twosixlabs.com, {friedman, ukuter}@sift.net

Abstract

The replicability of research is crucial for building trust in the peer review process and transitioning knowledge to real-world applications. While manual peer review excels in some regards, the variability of reviewer expertise, publication requirements, and research domains brings about uncertainty in the process. Replicability, in particular, is not necessarily a priority; this is evidenced by repeated failures in replication attempts such as the Psychology Reproducibility Project, where 61 of 100 replications fail. Improving human comprehension of decisive factors is crucial for integrating automated systems for replicability prediction into the review process. We develop a robust, automated method for semantic parsing, information extraction, and replication prediction that operates directly on PDFs. We introduce features that have not been explored in prior work, construct argument structures to guide understanding, and provide preliminary results for replication prediction.

1 Introduction

The replicability of research is crucial for building trust in the peer review process and for the transition of knowledge to real-world applications. Unfortunately, current attempts at replicating research show that many research papers do not replicate, with 61 of 100 failing the Psychology Reproducibility Project (Open Science Collaboration et al. 2015), 7 of 18 failing laboratory economics experiments (Camerer et al. 2016), 3 of 13 failing the Many Labs Replication Project (Klein et al. 2014), and more.

Currently, research is manually peer reviewed by a few experts often donating their time via venues such as conferences and journals. While manual peer review excels in some regards, the variability of reviewer expertise, publication requirements, and research domains brings about multiple levels of uncertainty. Additionally, peer review does not specifically attempt to identify the replicability of research, and, despite the increasing amount of automated analysis tools and replication prediction systems, there have been few changes to the review process over the years.

Determining replicability at review time is challenging for a multitude of reasons: limited access to data, limited reviewer time, inability to run new experiments, misleading statistics (Head et al. 2015), and the myriad variables that affect a reviewer's perception of the research, such as the readability of the explanations, clarity and detail of the methodology, significance of the authors' claims, etc. These variables that determine replicability can have varying levels of impact on the decision to accept a paper due to reviewer bias, research domain, and prior standards for acceptance. Not all acceptances of research are because it is replicable. Mapping these variables to actual replication outcomes can produce a less biased estimation of replicability.

In this work, we develop a novel method for understanding replicability given only a PDF of the research while encapsulating a wider, more robust set of factors than prior art. Using a combination of rule-based processing and machine learning, we perform consistent semantic parsing, feature extraction, and replicability classification.

Our main contributions are as follows:

- Consistent text extraction
- Automated classification of semantic flow
- Multifaceted feature extraction
- Preliminary replication prediction results

2 Related Work

The work related to our contributions are multiple-fold: previous literature has attempted to achieve a similar goal of predicting replicability, but there are a variety of methods that are relevant to our pipeline that have not been used for replicability prediction. We cover both aspects of that prior work here.

2.1 Predicting Replicability

The replication crisis is repeatedly noted throughout replicability literature (Open Science Collaboration et al. 2015; Klein et al. 2014; Camerer et al. 2016). Because peer review is currently an entirely manual process, a natural consequence is the desire to automate the understanding of replicability. An early attempt uses prediction markets to determine a ”market price” for research studies, representing the likelihood that those studies would replicate (Dreber et al. 2015). The prediction markets correctly predict 29/41 (71%) replications, but this method still requires approximately 50 domain experts to participate in the market. This is an impractical requirement for situations such as peer-reviewed conferences that often have three reviewers per paper. In (Altmejd et al. 2019), the authors attempt to predict the replicability of research by gathering features from within or about the research itself: this involves statistical design properties such as sample size, effect size, and p-value; or descriptive aspects, such as the number of citations, number of authors, and how subjects are compensated. By aggregating a dataset of 131 direct replications, they achieve approximately 70% prediction accuracy with random forest models. Although the feature extraction is still a manual process, locating relevant features in a paper is a tractable problem for an individual, which is a substantial improvement over the the prediction markets. (Yang, Youyou, and Uzzi 2020) take the automation a step further, and they obtain a 69% prediction accuracy by training on word embeddings of the research manuscript's text. We automatically extract features of prior work, generate a new set of features, and estimate replicability with higher accuracy.

2.2 Natural Language Processing

Whether attempting to obtain statistical test information or to operate directly on text, natural language processing is critical to the automation of manuscript featurization. A crucial innovation in the realm of general purpose natural language modeling is the use of models such as BERT (Devlin et al. 2018), which are pre-trained on large, unsupervised corpora. Through a fine-tuning step, these models transfer to new problems and domains. A particularly relevant application is SciBERT (Beltagy, Lo, and Cohan 2019), which is pre-trained on scientific publications from multiple domains. The authors show that this pre-training significantly improves results on downstream tasks related to scientific language. Recent work that focuses on scientific articles leverages these models to identify entities (Hakala and Pyysalo 2019), extract events and relationships (Allen et al. 2015; Valenzuela-Escárcega et al. 2018), and relate extracted events to domain models (Friedman et al. 2017). Our work utilizes fine-tuning to create span-based information extraction with a broader context that includes sample sizes, experimental methodologies, excluded sample counts, statistical tests, and more. Additionally, rather than focusing on the findings and contributions of scientific articles, we characterize methodologies, materials, confidence, and replicability.

3 Approach

We use a multi-stage pipeline in order to modularize each component of the extraction and prediction process. Each component can be easily changed out as enhancements to any components are developed, such as improvements in PDF parsing, new rules in rule-based methods, and updates for machine learning models. Figure 1 shows the flow of raw PDFs through various components, leading to the output of JSON files that are formatted with the article's text and associated features that can be used in downstream models. The pipeline comprises several main components: PDF extraction, semantic tagging, and information extraction.

3.1 PDF Extraction

Extracting text from a PDF and formatting it into informative segments are necessary steps for employing natural language approaches. We use Automator (Waldie 2009) to run the built-in PDF to RTF extraction tool. The RTF files maintain formatting information, but we use the command line utility *textutil* to convert the RTF files to HTML files, which we find to be more amenable to rule-based processing. We apply rules to the extraction because it is an erroneous process that fails around artifacts such as tables, captions, or footnotes. HTML representations are each parsed into a hash map where the keys are content styles and the values are all concatenated words and white-spaces of that style in the order they appear. The main content string of the paper is identified as the longest value, by character count, in this hash map. The main content string is used for all subsequent processing.

3.2 Semantic Tagging

A key element to understanding the structure of an argument is the semantic context in which the argument is made. To that end, we develop a machine learning model to annotate paragraphs based on their content. This is similar to the annotation work presented in (Chan et al. 2018), (Huber and Carenini 2019), and (Dasigi et al. 2017). Here though, we modify the annotation scheme to better match the problem of information extraction for replication prediction. We infer the discourse class for each sentence and perform an averaging of outputs to obtain the final class. This yields the following modified annotation scheme with 6 elements:

- **Introduction**: Problem statement and paper structure.
- **Methodology**: Specifics of the study, including participants, materials, and models.
- **Results**: Experimental results and statistical tests.
- **Discussion**: Author's interpretation of results and implications for the findings.
- **Research Practice**: Conflicts of interest, funding sources, and acknowledgements.
- **Reference**: Citations

Annotating Training Data In order to create a training set for discourse class prediction, we extract text from 838 social and behavioral science (SBS) research articles. In addition to the full text, these extractions contain the section header. This is what we use as our annotation, resulting in 81,001 labeled sentences. Due to the variation in section header names as a result of domain, tradition, or personal preference, we assign a set of keywords to each discourse class and label a section/segment of text if the section header is grammatically close to a keyword. The keywords used to create the dataset are:

- **Introduction**: {Introduction}
- **Methodology**: {Methodology, Analysis, Experiment, Method, Procedure, Design, Material, Participant}
- **Results**: {Results}
- **Discussion**: {Discussion, Conclusion}

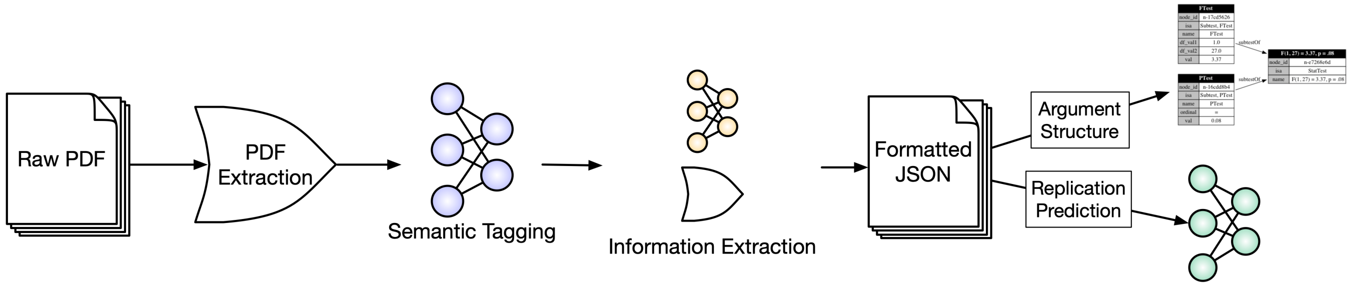


Figure 1: The full pipeline. We combine PDF extraction and rule-based parsing to generate strings of the research text, apply machine learning-based semantic tagging, and then extract features with machine-learning and rule-based approaches to generate a single formatted JSON per paper. This formatted JSON is convenient for downstream models, such as argument structure construction and replication prediction.

Table 1: The number of sentences per discourse tag extracted from the training data.

Discourse Tag	**Sentence Count**
Introduction	13,023
Methodology	24,930
Results	18,308
Discussion	14,233
Research Practices	353
Reference	10,153

- **Research Practices**: {Acknowledgements, Funding, Ethics Statement, Competing Interests, Ethical Approval}
- **Reference**: {Reference, Bibliography}

Creating Semantic Vector Representations Given a sentence extracted from a research article, we use the Universal Sentence Encoder model (Cer et al. 2018), which is designed to embed words, sentences, and small paragraphs into a semantically-related latent space. We represent a sentences as a 512-dimensional vector that encode the general semantics and context.

We use that 512-dimensional vector as input to a fully-connected hidden layer of size 512, followed by another full-connected hidden layer of size 256, followed by an output layer of size 6 (representing the discourse classes). A softmax activation after the output layer provides the discourse prediction. We use 50% dropout between the layers and a balanced sampling scheme to avoid overfitting to a single class. We use precision and recall to evaluate the prediction performance, shown in table 2.

The following is an example input whose actual section header is: **2.5 Inference** from (Cohan et al. 2020). Our model predicts the title: **Methodology**.

At inference time, the model receives one paper, P, and it outputs the SPECTER's Transformer pooled output activation as the paper representation for P (Equation 1). We note that for inference, SPECTER requires only the title and abstract of the given input paper; the model does not need any citation information about the input paper. This means that SPECTER can produce embeddings even for new papers that have yet to be cited, which is critical for applications that target recent scientific paper

Table 2: Precision/Recall/F1 results on a holdout set of annotated sentences.

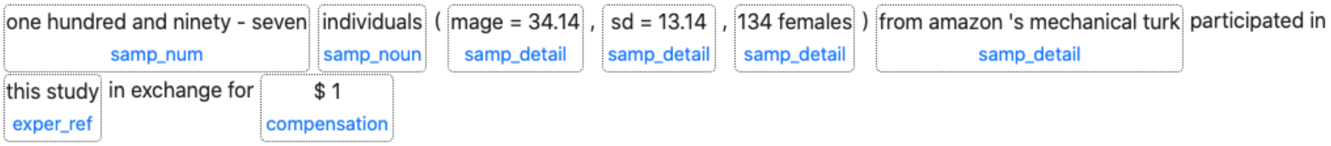
	Precision	Recall	F1-score
Introduction	0.53	0.70	0.60
Methodology	0.80	0.47	0.59
Results	0.56	0.64	0.60
Discussion	0.60	0.52	0.56
Research Practices	0.96	0.73	0.83
Reference	0.75	0.97	0.84

3.3 Information Extraction

In addition to the content and context extracted from the article, we further include features unrelated to the structure of the paper, but that are essential to the analysis of a paper's claims. These include both natural language features and statistical test results.

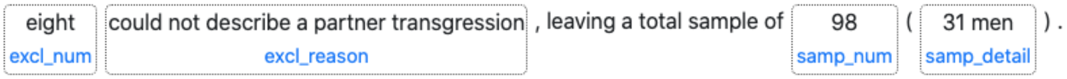
Language Quality Regardless of the validity of a paper's methodology and analysis, a failure to adequately communicate that information hinders others from using or replicating that research. As a means of assessing the quality of the writing itself, we compute three metrics over each paragraph in the text: readability, subjectivity, and sentiment. The idea for readability being related to the ability to reproduce findings is generated from the discussion in (Plavén-Sigray et al. 2017). We consider subjectivity due to discussions with social and behavioral science domain experts about inferring possible questionable research practices. Finally, positive or negative sentiment in the results or discussion sections may indicate biases towards the outcomes of the research. Although any one of these features may not directly express replicability, they do provide a holistic view of the writing.

Using each paragraph as input, we compute **readability** using Flesch Readability Ease (Kincaid et al. 1975), **sentiment** using the AllenNLP suite (Gardner et al. 2017), and **subjectivity** using the TextBlob package (Loria 2018). This produces a distribution of these features over the text that we



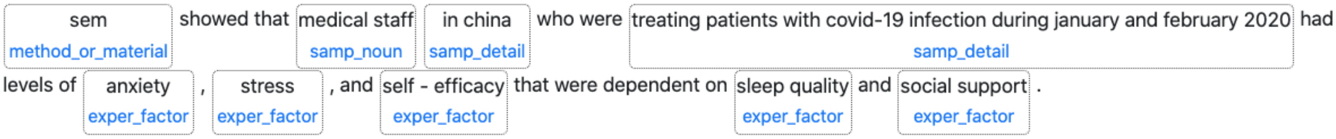
one hundred and ninety - seven individuals (mage = 34.14 , sd = 13.14 , 134 females) from amazon 's mechanical turk participated in this study in exchange for $ 1

Figure 2: Labeling spans for sample size (**samp_num**), sample details (**samp_detail**), and subject compensation (**compensation**) in a specific study (**exper_ref**).



eight could not describe a partner transgression , leaving a total sample of 98 (31 men) .

Figure 3: Labeling spans for the number of sample elements excluded (**excl_num**) and the stated reason they were excluded (**excl_reason**), as well as the final sample number.



sem showed that medical staff in china who were treating patients with covid-19 infection during january and february 2020 had levels of anxiety , stress , and self - efficacy that were dependent on sleep quality and social support .

Figure 4: Labeling the sample, experimental methods employed (**method_or_material**), and factors (**exper_factor**) under study.

can relate to discourse, experimental results (statistics), and domain-specific extractions.

3.4 Methodological Information Extraction

The unstructured prose of scientific documents includes key features for assessing replicability, such as sample sizes, populations, conditions, experimental variables, methods, materials, exclusion criteria, and participant compensation. Much of this information is available as concise spans of text in the document: *"twenty-four"* may be a sample size; *"undergraduates"* may be a population description; *"reaction time"* may be a dependent variable; and so on. Consequently, we are not interested in extracting and classifying *relations* at this phase of analyses; rather, we optimize our information extractor to classify individual *spans* within the text with context-sensitive labels.

Our dataset includes 620 labeled examples that are annotated with the following properties:

- **Sample count**: How many elements are in the sample.
- **Sample noun**: Noun phrases referring to sample elements, e.g., students, participants, cases, etc.
- **Sample detail**: Details of the same, e.g., race, sex, age, community, university, AMT, etc.
- **Compensation**: How participants are compensated.
- **Exclusion count**: Number excluded from sample.
- **Exclusion reason**: Stated reason(s) for why elements are excluded from the final sample set.
- **Experiment reference**: Name or reference to an experiment within the document.
- **Experimental condition**: Named or unnamed control or experimental condition employed.
- **Experimental variable/factor**: Elements measured or reported in the document, e.g., reaction time, participant preference, accuracy on a task.
- **Method or material**: experimental methods or materials employed, e.g., ANOVA, questionnaire, priming.

We extract these features using a transformer-based, token-level classifier that processes each sentence separately. The output of the classifier model is a Begin/Inside/Outside (BIO) prediction for each token in a sentence. This assumes that no labels overlap in the sentence, which is one constraint of our dataset.

We illustrate the above labels as predicted on some typical sentences from research articles in the SBS literature. Figure 2 shows our model's information extraction results for a typical statement introducing a population and sample size. This tags the English spans for sample count *"one hundred and ninety - seven,"* sample noun *"individuals,"* details (i.e., age mean, SD, gender, and AMT), an experiment reference to *"this study,"* and the compensation of *"$1."* In another paper, Figure 3 identifies the number of sample elements excluded, along with the resulting sample number and gender details. Finally, Figure 4 shows a sentence from the summary of an article, tagging *"sem"* (Structural Equation Modeling) as a methodology, sample noun and details, and five experimental factors that are assessed in the paper.

Our model next processes the resulting classified spans – as shown in Figures 2, 3, and 4 – to opportunistically extract domain-specific numerical and Boolean features. For example, the sample count and exclusion count are both expected to be integers, so it attempts to coerce *"one hundred and ninety - seven"* (Figure 2) and *"Eight"* (Figure 3) to integers and populate corresponding integer features. Similarly, the model uses a lexicon-based approach over the

Table 3: Precision/Recall/F1 results on a holdout set of information extraction examples.

Transformer Model	Precision	Recall	F1
distilbert_uncased	0.62	0.70	0.66
roberta_base	0.59	0.64	0.61
bert_large_uncased	0.61	0.71	0.66
scibert_scivocab_uncased	**0.67**	**0.74**	**0.70**
scibert_scivocab_cased	0.62	0.73	0.67

sample descriptor spans to populate Boolean features indicating whether participants' genders, age, race, religion, and community are specified, what the recruitment pool is (e.g., AMT, universities, etc.), and how they are compensated (e.g., course credit, monetary, etc.). These numerical, Boolean, and lexical features populate the *argument structure* of the paper, which we describe in subsequent sections.

We train a model by fine-tuning SciBERT and DistilBERT uncased models, and we evaluate using the same 558/62 randomized train/test split of our 620 labeled examples. Table 3 shows the results across four different transformer models for 100 iterations each, showing best performance from the SciBERT uncased model. While our model shows favorable results for our relatively small dataset of 620 examples, we are presently extending our dataset.

One limitation of the present sentence-level analysis is that cross-sentence coreferring expressions are unresolvable within the model, although – since we are not extracting complex relations across entities – most context-sensitive concepts such as sample-size and exclusion-count have ample context within the sentence itself. We plan to quantify the benefit of adding cross-sentence coreference resolution in future work.

3.5 Statistical Test Extraction

The descriptions of statistical tests in scientific documents are much more structured than descriptions of samples, methods, and factors. Consequently, our system uses Python regular expressions (rather than a transformer-based model) to extract statistical tests, motivated by processing speed and tailorability. Our regular expressions identify 25 different statistical tests and values, including p, R, R^2, d, F-tests, T-tests, mean, median, standard deviation, confidence intervals, odds ratios, non-significance, and more. These regular expressions were implemented for this system and were not reused from a previous system.

Our statistical test extractor then clusters extracted elements by proximity: Figure 5 shows two statistical tests ($F(1, 27) = 3.37$ and $p = .08$) that correspond to the same result. The system expresses each sub-test (e.g., F-test and p-test) as a separate sub-test leaf of the overall statistical test. Each sub-test describes distinctive features; for example, the p-test includes a value and an **ordinal** feature with the value "=" since the authors reported equality instead of "<" or ">," and the F-test includes two degrees of freedom and a value. Clustering these statistical tests into subgraphs helps identify duplicate reports of experimental results, and it provides context for downstream graphical analysis and machine learning.

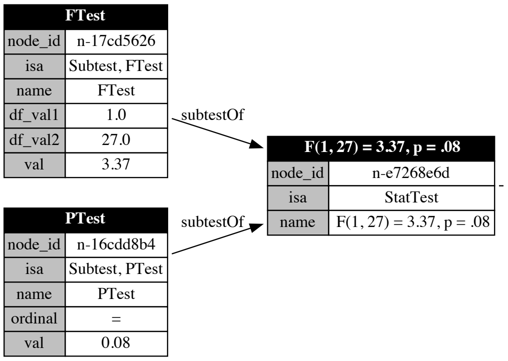


Figure 5: Semantic subgraph for a local cluster of two statistical tests extracted from a paper.

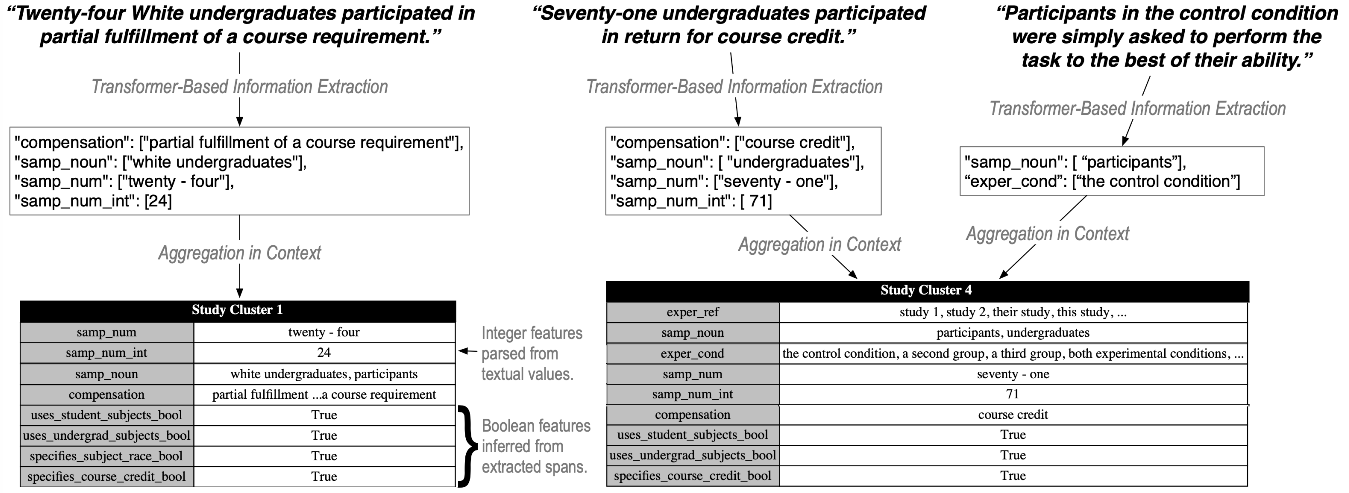
3.6 Assembly into Argument Structure

After extracting individual spans and subgraphs from the unstructured prose of a scientific article, we assemble the extracted information into a global graph that we refer to as the *argument structure* of the document. As implied by its name, the argument structure is designed to express the premises, evidence, and observations in a scientific article, ultimately in support of its conclusions.

The system generates the argument structure by iterating over the sequence of text segments and associated semantic tags (see Table 2 for a list of tags). Upon encountering a transition in semantic tags, such as a new **Methodology** section after a **Discussion** section, the system instantiates a new **Study** node within its argument structure, and then adds the BERT-extracted features (see above) and statistical test subgraphs (see above) as constituents of the new node. In this fashion, the system accumulates nodes for Introduction, Study, and Discussion prose. A small set of features for two Study nodes from the same paper are shown in Figure 6, populated by information extraction.

The graph-based layout of the argument structure allows the system to assess independent replicability concerns in a context-sensitive, explainable fashion. For example, as shown in Figure 6, the sample size of 24 for the study node at left may impact the judgment of that study's replicability, but it does not *necessarily* impact the replicability judgment of the study at right, in the same paper. Likewise specifying the participants' race in Figure 6 (left) may improve the replicability judgment of that study but should not affect the other study that does not specify participant race.

Each node in the directed argument structure graph is connected directly or indirectly to the node representing the scientific article itself. In this fashion, the argument structure is a fully-connected graph that supports graph and pattern matching, confidence propagation, and feature extraction in order to judge and explain replicability.



"Twenty-four White undergraduates participated in partial fulfillment of a course requirement."

```
"compensation": ["partial fulfillment of a course requirement"],
"samp_noun": ["white undergraduates"],
"samp_num": ["twenty - four"],
"samp_num_int": [24]
```

"Seventy-one undergraduates participated in return for course credit."

```
"compensation": ["course credit"],
"samp_noun": [ "undergraduates"],
"samp_num": ["seventy - one"],
"samp_num_int": [ 71]
```

"Participants in the control condition were simply asked to perform the task to the best of their ability."

```
"samp_noun": [ "participants"],
"exper_cond": ["the control condition"]
```

Study Cluster 1	
samp_num	twenty - four
samp_num_int	24
samp_noun	white undergraduates, participants
compensation	partial fulfillment ...a course requirement
uses_student_subjects_bool	True
uses_undergrad_subjects_bool	True
specifies_subject_race_bool	True
specifies_course_credit_bool	True

Study Cluster 4	
exper_ref	study 1, study 2, their study, this study, ...
samp_noun	participants, undergraduates
exper_cond	the control condition, a second group, a third group, both experimental conditions, ...
samp_num	seventy - one
samp_num_int	71
compensation	course credit
uses_student_subjects_bool	True
uses_undergrad_subjects_bool	True
specifies_course_credit_bool	True

Transformer-Based Information Extraction
Transformer-Based Information Extraction
Transformer-Based Information Extraction
Aggregation in Context
Aggregation in Context
Aggregation in Context
Integer features parsed from textual values.
Boolean features inferred from extracted spans.

Figure 6: Populating argument structure for two studies using information extracted across sentences and paragraphs.

3.7 Replicability Prediction

We train a random forest model to classify the replicability of of papers. Our dataset is a collection of papers from the Journal of Experimental Psychology, and we are able to mostly separate the replicable and non-replicable experiments. We plan to improve that separation and the calibration of the replicability scores in future work.

Ground Truth Replications To evaluate the ability of our model to correctly separate replicated studies from those that did not replicate, we train and test on replication attempts for the papers from the Journal of Experimental Psychology. We are able to collect approximately 150 PDFs of these papers for parsing and processing. As the replication studies are performed by different groups, there is variability in the number of features available in the given data. Many contain simple statistics such as sample size, but only a few contain p-value. To expand the set of available features, we manually mine them from the parsed PDFs. This gives features related to the number and significance of p-values reported, a proxy to the number of figures present, the presence of effect size, and the presence of an appendix. Furthermore, we judge replicability based on the percentage of known replications to known failures (e.g. in a set of replication studies, if an experiment was replicated 5 times and failed to replicate 3 times, we say the experiment replicated).

Prediction Model & Results We train a binary random forest classifier in a similar fashion to (Altmejd et al. 2019) to predict the replicability of an experiment. We use 5000 estimators with a max depth of 3. We evaluate our performance with AUC and accuracy, shown in Table 4. We select 11 psychology papers from the dataset and use these as the evaluation set. We predict using experiment p-value and the presence of effect size (binary). The results for the individual papers are shown in Table 5.

4 Discussion

Targeting replicability in the evaluation of research is a diverse task that is often not prioritized during peer review. Improving human comprehension of decisive factors is a crucial push towards integrating automated systems for replicability prediction into the review process. In this work, we develop an automated system for identifying, extracting, and organizing those factors. We introduce measures of language quality such as subjectivity, sentiment, and readability; we semantically tag text in order to understand language context; we extract statistical test information, linguistic relationships, and methodologies; and then we construct a hierarchical argument structure and perform replicability classification. These factors and their organization are intuitive to readers and allow for both top-down and bottom-up understanding of a paper's methods. Although leaving the review process entirely up to automation is not feasible, human-in-the-loop systems that guide reviewers through important text, factors, and predictions can reduce the amount of non-replicable papers that make it through review.

5 Future Work

One of the main focuses of our future work is to extend our ground truth datasets and evaluate replicability prediction across the combinations of features that we develop in the current work. Due to the limited data size, the evaluation set is too small to definitively select the best combination of features for replicability prediction.

We are also working to broaden our system's features and capabilities. For instance, we are incorporating a transformer-based information extractor that extracts the causal, proportional, and comparative relationships in scientific claims (Magnusson and Friedman 2021) to relate the claims within and across scientific documents in our corpus. To improve human interpretation, we are working to produce an explainability interface for users to inspect our extractions, predictions, and argument structure for guided paper understanding.

Table 4: The accuracy and AUC for a random forest classifier with 5000 estimators and a max depth of 3.

	Accuracy	AUC
Evaluation Set	0.90	0.89

Table 5: The individual predictions and labels for each paper in the evaluation set. The model correctly predicts 10 of the 11 papers.

Paper Reference	Label	Prediction
(Nosek, Banaji, and Greenwald 2002) Exp. 1	1	0.70
(Nosek, Banaji, and Greenwald 2002) Exp. 2	1	0.66
(Soto et al. 2008)	1	0.66
(Monin, Sawyer, and Marquez 2008)	0	0.36
(Purdie-Vaughns et al. 2008)	0	0.31
(Goff, Steele, and Davies 2008)	0	0.27
(Payne, Burkley, and Stokes 2008)	1	0.27
(Shnabel and Nadler 2008)	0	0.25
(Lemay Jr and Clark 2008a)	0	0.24
(Fischer, Greitemeyer, and Frey 2008)	0	0.23
(Lemay Jr and Clark 2008b)	0	0.06

We will further assess the validity of the elements of a paper, i.e., the confidence that we have in the claims in the paper given the assumptions made by it, by using an existing probabilistic inference and recognition system, SUNNY, originally developed for planning (Kuter et al. 2004) and social network analysis (Kuter and Golbeck 2007, 2010). SUNNY propagates local estimates of uncertainty through large models. Its most basic output is the probability, as a function of time, that a particular event will be true. We have extended SUNNY for k-nearest neighbors (kNN) learning and prediction capabilities, as well as a Naive Bayes diagnoses of confidence scores based on the (kNN) clustering. The numeric and qualitative features in our argument structures form the basis of the kNN clustering and we will extend these measures towards predicting replicability scores in SUNNY in the near future.

Acknowledgements

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) and Army Research Office (ARO) under Contract No. W911NF-20-C-0002. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA) and Army Research Office (ARO).

References

Allen, J.; de Beaumont, W.; Galescu, L.; and Teng, C. M. 2015. Complex event extraction using DRUM. Technical report, Florida Institute for Human and Machine Cognition Pensacola United States.

Altmejd, A.; Dreber, A.; Forsell, E.; Huber, J.; Imai, T.; Johannesson, M.; Kirchler, M.; Nave, G.; and Camerer, C. 2019. Predicting the replicability of social science lab experiments. *PloS one* 14(12).

Beltagy, I.; Lo, K.; and Cohan, A. 2019. SciBERT: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676* .

Camerer, C. F.; Dreber, A.; Forsell, E.; Ho, T.-H.; Huber, J.; Johannesson, M.; Kirchler, M.; Almenberg, J.; Altmejd, A.; Chan, T.; et al. 2016. Evaluating replicability of laboratory experiments in economics. *Science* 351(6280): 1433–1436.

Cer, D.; Yang, Y.; Kong, S.-y.; Hua, N.; Limtiaco, N.; John, R. S.; Constant, N.; Guajardo-Cespedes, M.; Yuan, S.; Tar, C.; et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175* .

Chan, J.; Chang, J. C.; Hope, T.; Shahaf, D.; and Kittur, A. 2018. Solvent: A mixed initiative system for finding analogies between research papers. *Proceedings of the ACM on Human-Computer Interaction* 2(CSCW): 1–21.

Cohan, A.; Feldman, S.; Beltagy, I.; Downey, D.; and Weld, D. S. 2020. Specter: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2270–2282.

Dasigi, P.; Burns, G. A.; Hovy, E.; and de Waard, A. 2017. Experiment segmentation in scientific discourse as clause-level structured prediction using recurrent neural networks. *arXiv preprint arXiv:1702.05398* .

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

Dreber, A.; Pfeiffer, T.; Almenberg, J.; Isaksson, S.; Wilson, B.; Chen, Y.; Nosek, B. A.; and Johannesson, M. 2015. Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences* 112(50): 15343–15347.

Fischer, P.; Greitemeyer, T.; and Frey, D. 2008. Self-regulation and selective exposure: the impact of depleted self-regulation resources on confirmatory information processing. *Journal of personality and social psychology* 94(3): 382.

Friedman, S.; Burstein, M.; McDonald, D.; Plotnick, A.; Bobrow, L.; Bobrow, R.; Cochran, B.; and Pustejovsky, J. 2017. Learning by reading: Extending and localizing against a model. *Advances in Cognitive Systems* 5: 77–96.

Gardner, M.; Grus, J.; Neumann, M.; Tafjord, O.; Dasigi, P.; Liu, N. F.; Peters, M.; Schmitz, M.; and Zettlemoyer, L. S. 2017. AllenNLP: A Deep Semantic Natural Language Processing Platform.

Goff, P. A.; Steele, C. M.; and Davies, P. G. 2008. The space between us: stereotype threat and distance in interracial contexts. *Journal of personality and social psychology* 94(1): 91.

Hakala, K.; and Pyysalo, S. 2019. Biomedical named entity recognition with multilingual BERT. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, 56–61.

Head, M. L.; Holman, L.; Lanfear, R.; Kahn, A. T.; and Jennions, M. D. 2015. The extent and consequences of p-hacking in science. *PLoS Biol* 13(3): e1002106.

Huber, P.; and Carenini, G. 2019. Predicting discourse structure using distant supervision from sentiment. *arXiv preprint arXiv:1910.14176* .

Kincaid, J. P.; Fishburne Jr, R. P.; Rogers, R. L.; and Chissom, B. S. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.

Klein, R. A.; Ratliff, K. A.; Vianello, M.; Adams Jr, R. B.; Bahník, Š.; Bernstein, M. J.; Bocian, K.; Brandt, M. J.; Brooks, B.; Brumbaugh, C. C.; et al. 2014. Investigating variation in replicability. *Social psychology* .

Kuter, U.; and Golbeck, J. 2007. Sunny: A new algorithm for trust inference in social networks using probabilistic confidence models. In *AAAI*.

Kuter, U.; and Golbeck, J. 2010. Using Probabilistic Confidence Models for Trust Inference in Web-Based Social Networks. *Transactions on Internet Technology (TOIT)* 7: 1377–1382.

Kuter, U.; Nau, D.; Gossink, D.; and Lemmer, J. F. 2004. Interactive Course-of-Action Planning Using Causal Models. In *International Conference on Knowledge Systems for Coalition Operations (KSCO-2004)*, 37–52.

Lemay Jr, E. P.; and Clark, M. S. 2008a. " Walking on eggshells": how expressing relationship insecurities perpetuates them. *Journal of personality and social psychology* 95(2): 420.

Lemay Jr, E. P.; and Clark, M. S. 2008b. How the head liberates the heart: Projection of communal responsiveness guides relationship promotion. *Journal of Personality and Social Psychology* 94(4): 647.

Loria, S. 2018. textblob Documentation. *Release 0.15* 2.

Magnusson, I. H.; and Friedman, S. E. 2021. Graph Knowledge Extraction of Causal, Comparative, Predictive, and Proportional Associations in Scientific Claims with a Transformer-Based Model. In *AAAI Workshop on Scientific Document Understanding*.

Monin, B.; Sawyer, P. J.; and Marquez, M. J. 2008. The rejection of moral rebels: resenting those who do the right thing. *Journal of personality and social psychology* 95(1): 76.

Nosek, B. A.; Banaji, M. R.; and Greenwald, A. G. 2002. Math= male, me= female, therefore math$\neq$ me. *Journal of personality and social psychology* 83(1): 44.

Open Science Collaboration; et al. 2015. Estimating the reproducibility of psychological science. *Science* 349(6251).

Payne, B. K.; Burkley, M. A.; and Stokes, M. B. 2008. Why do implicit and explicit attitude tests diverge? The role of structural fit. *Journal of personality and social psychology* 94(1): 16.

Plavén-Sigray, P.; Matheson, G. J.; Schiffler, B. C.; and Thompson, W. H. 2017. The readability of scientific texts is decreasing over time. *Elife* 6: e27725.

Purdie-Vaughns, V.; Steele, C. M.; Davies, P. G.; Ditlmann, R.; and Crosby, J. R. 2008. Social identity contingencies: how diversity cues signal threat or safety for African Americans in mainstream institutions. *Journal of personality and social psychology* 94(4): 615.

Shnabel, N.; and Nadler, A. 2008. A needs-based model of reconciliation: satisfying the differential emotional needs of victim and perpetrator as a key to promoting reconciliation. *Journal of personality and social psychology* 94(1): 116.

Soto, C. J.; John, O. P.; Gosling, S. D.; and Potter, J. 2008. The developmental psychometrics of big five self-reports: Acquiescence, factor structure, coherence, and differentiation from ages 10 to 20. *Journal of personality and social psychology* 94(4): 718.

Valenzuela-Escárcega, M. A.; Babur, Ö.; Hahn-Powell, G.; Bell, D.; Hicks, T.; Noriega-Atala, E.; Wang, X.; Surdeanu, M.; Demir, E.; and Morrison, C. T. 2018. Large-scale automated machine reading discovers new cancer-driving mechanisms. *Database* 2018.

Waldie, B. 2009. Automator for Mac OS X 10.6 Snow Leopard: Visual QuickStart Guide .

Yang, Y.; Youyou, W.; and Uzzi, B. 2020. Estimating the deep replicability of scientific findings using human and artificial intelligence. *Proceedings of the National Academy of Sciences* 117(20): 10762–10768.