Research Replication Prediction Using Weakly Supervised Learning

Tianyi Luo, Xingyu Li, Hainan Wang, Yang Liu
Computer Science and Engineering, University of California, Santa Cruz, CA
{tluo6, xli279, hwang207, yangliu}@ucsc.edu

Abstract

Knowing whether a published research result can be replicated is important. Carrying out direct replication of published research incurs a high cost. There are efforts tried to use machine learning aided methods to predict scientific claims' replicability. However, existing machine learning aided approaches use only hand-extracted statistics features such as p-value, sample size, etc. without utilizing research papers' text information and train only on a very small size of annotated data without making the most use of a large number of unlabeled articles. Therefore, it is desirable to develop effective machine learning aided automatic methods which can automatically extract text information as features so that we can benefit from Natural Language Processing techniques. Besides, we aim for an approach that benefits from both labeled and the large number of unlabeled data. In this paper, we propose two weakly supervised learning approaches that use automatically extracted text information of research papers to improve the prediction accuracy of research replication using both labeled and unlabeled datasets. Our experiments over real-world datasets show that our approaches obtain much better prediction performance compared to the supervised models utilizing only statistic features and a small size of labeled dataset. Further, we are able to achieve an accuracy of $\mathbf{75.76}\%$ for predicting the replicability of research.

1 Introduction

Non-reproducible scientific results will mislead the progress of science and undermine the trustworthiness of the research community. In recent years, we saw the emergence of systematic large-scale replication projects which are based on the concerns of research credibility in the social and behavioral sciences (Camerer et al., 2016, 2018; Ebersole et al., 2016; Klein et al., 2014b, 2018; Collaboration et al., 2015). Researchers conducted preregistered replications of hundreds of classic and contemporary published findings in the social and behavioral sciences. Unfortunately, the reported replication rates only range from 39% to 62%. Therefore it is important to develop a confidence scoring system for the following question:

To what extend can a research result be reproduced?

The answer to the above question will help facilitate the policymakers as well as the general public to better understand and digest a published claim. As a response, for example, Defense Advanced Research Projects Agency (DARPA) has announced a systematic confidence checking of published claims (Russell, 2019).

Alongside the above encouraging movement, the downside is that the average replication expense of each research project (which often consists a number of research studies) can go up to $500,000 (Freedman et al., 2015)[1], which is hardly affordable to replicate each research finding, with an exponentially increasing number of publications.

Recently, efforts have been noted to use machine learning as a much cheaper and more efficient alternative to provide an informative replication prediction (Dreber et al., 2019; Yang, 2018; Altmejd et al., 2019). It has been reported that with simple machine learning models, a predicted accuracy of 71% can be achieved. Although we should not trust or rely on a machine-made prediction entirely, such automatic predictions offer cheap, scalable, and useful information for performing targeted spot-checking and for raising a red flag towards a partic-

[1] "Irreproducibility also has downstream impacts in the drug development pipeline. Academic research studies with potential clinical applications are typically replicated within the pharmaceutical industry before clinical studies are begun, with each study replication requiring between 3 and 24 months and between US$500,000 to US$2,000,000 investment"

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 1464–1474

ular scientific claim.

Nonetheless, existing machine learning works on replication prediction face a couple of outstanding challenges:

- Substantial human efforts are required to extract features from the published articles, such as p-values of the claims, effect size, author information, etc. to train a supervised machine learning model;
- The small amount of expensive annotated training data will limit the use of more sophisticated but more accurate learning techniques (e.g., deep neural networks based natural language processing tools).

We aim for a method that is fully automatic in feature generation, and that can leverage the existence of the large corpus of unlabeled (checked) articles for boosting up the performance in predicting replications.

To tackle the first challenge, we will resort to natural language processing (NLP) tools to process the research articles to obtain meaningful text features. Text information of research papers is important and intuitive resource for training machine learning models. The rich amount of structural text information looks promising to us to help improve the predictive performance of replication. Further, a good understanding of text information from different components of an article (e.g., abstract, introduction, methods, experimental results, etc.) will also be helpful for highlighting suspicious sections of the articles for a more targeted check.

However, the training of the state-of-the-art NLP models aligns with our second challenge that it often relies on a massive volume of annotated training data. Due to the severely limited ground truth annotation we have, we desire a method that leverages large amounts of unlabeled research articles. These unlabeled examples, although possibly noisy, can provide informative features.

To make the most use of the unlabeled data, we explore the possibility of using a weakly supervised approach to perform replication prediction. The particular type of weakly supervised learning method that we will focus on utilizes techniques from the literature on learning from noisy labels (Liu et al., 2012; Natarajan et al., 2013; Scott, 2015; Van Rooyen et al., 2015; Liu and Guo, 2020). Our high-level idea is to bootstrap the small set of labeled data to train a set of weak predictors which will help us generate "artificial" and noisy labels for the unlabeled articles. Then we will apply tools from learning with noisy labels to improve the training with these artificially supervised examples.

We focus on two approaches to address the above problem of learning with artificial labels. The first approach uses efficient variational inference methods (Liu et al., 2012) to estimate the error rates of the noisy labels. The above knowledge of error rates allows us to perform loss correction (Natarajan et al., 2013) to improve the performance with the help of an unlabeled dataset. The second approach is inspired by a recent work (Liu and Guo, 2020) that proposed a family of peer loss functions which can perform learning with noisy labels without knowing noise rates and without conducting intermediate error rates' estimation step.

We utilized both labeled and unlabeled datasets to carry out the study of replication prediction. The labeled dataset containing 399 research articles are obtained from summarizing eight research replication projects (details will be given later). As for the unlabeled dataset, a python crawler is implemented to obtain the pdf files of 2,170 research papers from the websites of corresponding journals. We preprocess the files to extract text information. Then BERT (Devlin et al., 2018) is used for tokenization and for obtaining word embeddings to serve as the input features for training.

The experimental results demonstrate that i) using text information as features can improve the performance than utilizing only pre- and hand-extracted statistics features. The combination of models trained on text features and statistics features separately can obtain better performance than separate models; and ii) our weakly supervised methods that take advantage of unlabeled data can significantly improve the prediction performance. The best of our proposed methods can achieve a prediction accuracy of $\mathbf{75.76}\%$, as well as a $\mathbf{72.50}\%$ precision, a $\mathbf{88.24}\%$ recall, and a $\mathbf{78.95}\%$ F1 score.

We summarize our contributions as follows: (1) We propose two weakly supervised learning approaches based on text information of research papers to improve the prediction accuracy of research replication using both labeled and unlabeled datasets. (2) We present experimental results to validate the usefulness of our proposed weakly supervised learning models. (3) We contribute to the community by

publishing our codes and data. Please refer to https://github.com/pkuluotianyi/PeerRRP for the most updated codes and datasets.

2 Related Work

Replication crisis has spurred systematic large-scale direct replication projects in the social and behavioral sciences (Camerer et al., 2016, 2018; Ebersole et al., 2016; Klein et al., 2014b, 2018; Collaboration et al., 2015). Data is collected by individual volunteers, volunteer teams, or Amazon Mechanical Turk (AMT). However, direct research replication is expensive and time-consuming (Freedman et al., 2015). Machine learning serves as a much more efficient method to conduct replication prediction. Altmejd et al. (2019) applied ML methods on the data from four large-scale replication projects in experimental psychology and economics and studied which variables drive predictable replication. But they used only statistics features such as p-value, sample size, etc. and train only on a small labeled dataset.

We hold the hypothesis that text features contain rich information to potentially improve the performance of replication prediction. In NLP, many research works have been proposed for text processing to make use of text features (Jurafsky and Martin, 2014; Biemann and Mehler, 2014; Boroş et al., 2018; Devlin et al., 2018).

Weakly supervised learning approaches have been proposed to utilize both labeled and unlabeled data (Zhou, 2018; Oliver et al., 2018; Miyato et al., 2018). Our weakly supervised learning approaches tie close to learning with the inaccurate supervision (Cesa-Bianchi et al., 2011; Bylander, 1994; Scott et al., 2013; Scott, 2015; Van Rooyen et al., 2015). Particularly relevant to us, a surrogate loss function is proposed in (Natarajan et al., 2013) to achieve an unbiased estimation of the true training loss using only noisy labels. Liu and Guo (2020) introduced a new family of loss functions, peer loss functions, to empirical risk minimization (ERM), for a broad class of learning with noisy labels problems, without requiring estimating the error rates of the noisy labels.

3 Datasets

Annotated Data In our study, we obtained 399 annotated articles containing labels indicating whether the involved research claim can be reproduced or not. If it can be replicated, we use the label '1' to denote it. Otherwise, the label '0' is used to represent it. There exist different definitions and criteria for a claim to be replicable. Here for the collected dataset, a claim extracted from the article is replicable if an independent effort can produce a statistically significant effect in the original direction as originally claimed.

The question of how we treat an article/claim as replicable is an active research question itself (Simonsohn, 2015). To include as many annotated data points as possible, we adopt the most basic binary model that defines replication success as a "statistically significant (p-value $<= 0.05$) effect in the same direction as in the original study." (Altmejd et al., 2019)

The annotated dataset comes from eight research replication projects which are the Registered Replication Report (RRR) (Simons et al., 2014), Many Labs 1 (Klein et al., 2014a), Many Labs 2 (Klein et al., 2018), Many Labs 3 (Ebersole et al., 2016), Social Sciences Replication Project (SSRP) (Camerer et al., 2018), PsychFileDrawer (Pashler et al., 2019), Experimental Economics Replication Project (Camerer et al., 2016), and Reproducibility Project: Psychology (RPP) (Collaboration, 2012).

Year	2011	2012	2013	2014	Total
# of pub	240	267	243	231	981

Table 1: Distribution of published economic related papers' number by year in the unlabeled dataset

Among 399 annotated samples, 201 samples are labeled as '1' (replicable). The remaining 198 samples are annotated as '0' (non-replicable). From the distribution of class labels, we observe that this annotated dataset is balanced.

Unsupervised Data In addition, we deployed a crawler to obtain an unlabeled dataset to pair with the above annotated one. Because the published research papers in the labeled dataset are mainly from American Economic Review and Psychological Science and all the other papers in the annotated dataset are economic and psychology-related, we use the crawler to get all 2,170 published research papers from the websites of American Economic Review (Jan 2011 - Dec 2014) and Psychological Science (Jan 2006 - Dec 2012) to form our unlabeled dataset. The number of papers crawled in the American Economic Review website is 981 and there are 1,189 papers from the Psychological Science website. The distribution of papers' number

Year	2006	2007	2008	2009	2010	2011	2012	Total
# of pub	185	200	196	238	293	243	224	1189

Table 2: Distribution of published psychological related papers' number by year in the unlabeled dataset

by year about American Economic review and Psychological Science are shown in Table 1 and Table 2 respectively.

Our setting is severely imbalanced: we have a very small amount of labeled data and a much large amount of unlabeled ones.

Datasets	# of docs	Avg len	Max len	Min len
Train	300	8948	68998	1446
Test	99	8343	33354	3599
Unlabeled	2170	6647	28994	1260

Table 3: Number, average length, maximum length, and minimum length of documents in different datasets

We list the average length (# of words contained), minimum length, and maximum length information of different datasets in Table 3.

4 Weakly Supervised Research Replication Prediction

We introduce the pipeline of our weakly supervised research prediction framework.

Feature Extraction Our method relies on automatically extracted text features. Specifically, PDFMiner (Shinyama, 2014) is used to extract the text information in the raw pdf files of the articles. Tf-idf features are used in bag-of-words models. BERT (Devlin et al., 2018) is used for tokenization and obtaining word embeddings as the input features of the sequential models. More specifically, we use "bert-base-uncased" pretrained model from Transformers (Wolf et al., 2019) which has 12-layer, 768-hidden, 12-heads, 110M parameters and trained on lower-cased English text.

Artificial and Noisy Label Generation Our problem is formulated as a binary classification to predict whether a research paper can be replicated or not. We utilize five basic classifiers trained on the labeled dataset to obtain artificial labels for the unlabeled articles. They are five commonly used binary classification algorithms including Logistic Regression (LR) (Peng et al., 2002), Random Forest (RF) (Ho, 1995), Support Vector Machine (SVM) (Chang and Lin, 2011), Multilayer Perceptron (MLP) (Goodfellow et al., 2016), and Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997).

Suppose that we have an annotated training dataset $\mathcal{L} := \{(x_i, y_i)\}_{i=1}^{L}$, an unlabeled dataset $\mathcal{U} := \{x_i\}_{i=1}^{U}$, and a test dataset $\mathcal{T} := \{(x_i, y_i)\}_{i=1}^{T}$, where $x_i \in X \subseteq R^d$ is a d-dimensional vector. We have K baseline classifiers $\mathcal{F} := \{f_1, f_2, ..., f_K : X \to \{0, 1\}\}$ that map each feature vector to a binary classification outcome. We let $N = L+U$, i.e., the total number of training dataset is N.

Given the whole training data $\mathcal{D} = \mathcal{L} \cup \mathcal{U}$ and multiple classifiers $\{f_j\}_{j=1}^{K}$, we firstly train five basic classifiers and get their predictions in $\mathcal{D} := \{(x_i, \bar{y}_i^j)\}_{i=1}^{N}$, $j = 1, ..., K$. Then we can use aggregation rules, e.g., majority voting rule, to obtain the noisy labels for the whole training data $\mathcal{Y}^{noise} := \{\bar{y}_i^{noise}\}_{i=1}^{N}$.

Training with Artificially Generated Noisy Labels Then we can utilize two different ways to conduct the learning with noisy labels $\mathcal{Y}^{noise}$. Details will be given in the next Section.

5 Method

In this section we present two weakly supervised methods. The first approach is based on the error correction proxy loss function (Natarajan et al., 2013) and the variational inference approaches (mean field) (Liu et al., 2012) to estimate the error rates. The two techniques jointly provide us a bias-corrected training process to improve the model's robustness against noises in labels. We name this solution as *Variational Inference aided Weakly Supervised Learning*.

The second approach is built on the peer loss approach (Liu and Guo, 2020). This approach is particularly suitable for our application when the label noises are unclear. In this paper, we will apply peer loss function in the weakly supervised learning scenario for the research replication prediction problem. We name this solution as *Peer Loss aided Weakly Supervised Learning*.

5.1 Variational Inference aided Weakly Supervised Learning

Algorithm 1 Variational Inference aided Weakly Supervised Learning

Require:

Input:

$\mathcal{D} = \{(x_1, y_1), ..., (x_N, y_N)\}$: training data

$\mathcal{L} = \{(x_1, y_1), ..., (x_L, y_L)\}$: labeled data

$\mathcal{U} = \{x_1, ..., x_U\}$: unlabeled data

$\mathcal{T} = \{(x_1, y_1), ..., (x_T, y_T)\}$: test data

$\mathcal{F} = \{f_1, ..., f_K\}$: classifiers

Ensure:

1: Train K classifiers ($\mathcal{F}$) on the labeled training data $\mathcal{L}$.
2: **for** $j = 1$ to K **do**
3: **for** $i = 1$ to N **do**
4: Compute $\bar{y}_i^j$ using j-th basic classifier.
5: **end for**
6: **end for**
7: Aggregate above labels into $\{\bar{y}_i^{noise}\}_{i=1}^N$ and estimate the error rates according to mean field method described in (Liu et al., 2012).
8: Train the LSTM model using the proxy loss function mentioned in Section 5.1 with the estimated error rates in line#7 as the inputs.
9: **for** $t = 1$ to T **do**
10: Output prediction.
11: **end for**

We start with using the five basic classifiers (LR, RF, SVM, MLP, and LSTM) trained on the annotated dataset of small size to generate the noisy labels for the whole training data respectively. These noisy labels will then be aggregated using a variational procedure (Liu et al., 2012), which we reproduce below:

Denote by μ_i as the probability of different class labels for the i-th train sample, ω_j as the weight or ability of the j-th classifier, α and β are the hyperparameters, $\delta_{ij} = 1[\bar{y}_i^j = \bar{y}_i^{noise}]$, and g is a function to calculate the error rates using $\{\bar{y}_i^{em}\}_{i=1}^N, \bar{\omega}_j$. μ_i and ω_j are firstly estimated using the Expectation-Maximization (EM) algorithms. We then obtain EM predictions $\bar{y}_i^{em}$ based on the above estimated μ_i and ω_j. $\bar{y}_i^{em}$ at the final step will serve as our noisy label $\bar{y}_i^{noise}$. The final step is to estimate error rates

$$\sigma_0 := P(\bar{y}_i^{noise} = 1 | y_i = 0)$$

and

$$\sigma_1 := P(\bar{y}_i^{noise} = 0 | y_i = 1)$$

by using $\bar{y}_i^{em}$ as the proxy for the ground truth label. The procedure is summarized in Algorithm 2. More detailed explanation are described in (Liu et al., 2012).

Algorithm 2 Aggregation and Error Rates

1: Update μ_i :

$$\mu_i(z_i) = \prod_{j \in K} \bar{\omega}_j^{\delta_{ij}} (1 - \bar{\omega}_j)^{1 - \delta_{ij}}$$

2: Update $\bar{\omega}_j$: $\bar{\omega}_j = \frac{\sum_{i \in N} \mu_i(\bar{y}_i^j) + \alpha}{N + \alpha + \beta}$
3: EM Predictions : $\bar{y}_i^{em} = \text{argmax}_z \ \mu_i(z_i)$
4: Error rates :

$$\sigma_0 = \frac{|i : \bar{y}_i^{em} = 0, \bar{y}_i^{noise} = 1|}{|i : \bar{y}_i^{em} = 0|}$$

$$\sigma_1 = \frac{|i : \bar{y}_i^{em} = 1, \bar{y}_i^{noise} = 0|}{|i : \bar{y}_i^{em} = 1|}$$

Finally, we use an LSTM neural network model with proxy loss function as shown in (Natarajan et al., 2013) to conduct the training. The definition of proxy loss function is as follows:

$$\sum_{i=1}^{N} \frac{(1 - \sigma_{1-y_i^p}) \ell(y_i^p, \bar{y}_i^{noise}) - \sigma_{y_i^p} \ell(1 - y_i^p, \bar{y}_i^{noise})}{1 - \sigma_1 - \sigma_0},$$

where in above $\ell(y_i^p, \bar{y}_i^{noise})$ is a standard cross entropy loss function where y_i^p is the i-th sample's real-value prediction of final LSTM model and $\bar{y}_i^{noise}$ is the corresponding noisy label.

The procedure is summarized in Algorithm 1.

5.2 Peer Loss aided Weakly Supervised Learning

Variational inference (VI) aided weakly supervised learning method requires estimating the error rates. This additional step of estimation may introduce estimation errors that can affect the final model's performance. Liu and Guo (2020) provided an alternative, peer loss, to deal with noisy labels that does not require an additional estimation step for the noise rates. We propose peer loss (PL) aided weakly supervised learning method.

Similar to the VI approach, we firstly train five basic classifiers on the annotated dataset of small size to provide the noisy supervisions for the whole training data $\mathcal{Y}^{noise} := \{\bar{y}_i^{noise}\}_{i=1}^N$, as mentioned in Section 4 via a simple majority vote.

For each training sample $(x_i, \bar{y}_i^{noise})$, we randomly draw another two samples

Peer Samples: $(x_{i_1^p}, \bar{y}_{i_1^p}^{noise}), (x_{i_2^p}, \bar{y}_{i_2^p}^{noise})$

such that $i_1^p \neq i_2^p$ and $i_1^p, i_2^p \neq i$. $(x_{i_1^p}, \bar{y}_{i_1^p}^{noise}), (x_{i_2^p}, \bar{y}_{i_2^p}^{noise})$ are the i-th data's peer samples. Then we calculate peer loss function as shown in (Liu and Guo, 2020). The definition of total peer loss $L_{peer}(\mathcal{Y}^p, \mathcal{Y}^{noise})$ is given as follows:

$$\sum_{i=1}^{N} \ell(y_i^p, \bar{y}_i^{noise}) - \alpha \cdot \ell(y_{i_1^p}^p, \bar{y}_{i_2^p}^{noise})$$

where $\ell(y_i^p, \bar{y}_i^{noise})$ is a standard cross entropy loss function where y_i^p is the i-th sample's real-value prediction of final LSTM model and $\bar{y}_i^{noise}$ is the corresponding noisy label. α is a hyperparameter that we will tune with.

We use an LSTM neural network model with the above defined peer loss function and train the model. The procedure is further illustrated in Algorithm 3.

Algorithm 3 Peer Loss aided Weakly Supervised Learning

Require:

Input:
$\mathcal{D} = \{(x_1, y_1), ..., (x_N, y_N)\}$: training data
$\mathcal{L} = \{(x_1, y_1), ..., (x_L, y_L)\}$: labeled data
$\mathcal{U} = \{x_1, , ..., x_U\}$: unlabeled data
$\mathcal{T} = \{(x_1, y_1), ..., (x_T, y_T)\}$: test data
$\mathcal{F} = \{f_1, ..., f_K\}$: classifiers

Ensure:

Train K classifiers ($\mathcal{F}$) on the labeled training data $\mathcal{L}$.
for $j = 1$ to K **do**
  for $i = 1$ to N **do**
    Compute $\bar{y}_i^j$ using j-th basic classifier.
  end for
end for
Compute $\{\bar{y}_i^{noise}\}_{i=1}^N$ using majority rule.
for $i = 1$ to N **do**
  Construct $\{(x_i, \bar{y}_i^{noise}), (x_{i_1^p}, \bar{y}_{i_2^p}^{noise})\}$.
end for
Create noisy training dataset: $\mathcal{D}^{noise} = \{(x_i, \bar{y}_i^{noise}), (x_{i_1^p}, \bar{y}_{i_2^p}^{noise})\}_{i=1}^N$.
Train the LSTM model using peer loss function as shown in Section 5.2 on $\mathcal{D}^{noise}$.
for $t = 1$ to T **do**
  Output prediction.
end for

5.3 Other Methods

To complete our analysis, we also take an off-the-shelf semi-supervised learning technique DIVIDEMIX (Li et al., 2020). It is a broad literature of methods proposed in semi-supervised learning and we chose the most recent and robust approach. DIVIDEMIX is a semi-supervised method which trains two networks simultaneously and the training dataset is dynamically divided into a labeled dataset and an unlabeled dataset in each iteration. We adapt the setting of DIVIDEMIX to ours to serve as a baseline comparison. DIVIDEMIX can benefit from the unlabeled data but they do not use bias-corrected loss function which is different from our methodology.

6 Experiments

In this section, we present our experimental results and findings and offer discussions.

6.1 Experimental Setup

We have 399 labeled and 2,170 unlabeled samples. Randomly selected 300 (150:1;150:0) labeled and 2,170 unlabeled samples are considered as the training dataset. We test our proposed framework on the remaining 99 (51:1;48:0) labeled replication projects.

We consider both text and statistics features of research papers. p-value, effect size, sample size are utilized as statistics features. As for the text information, Tf-idf and word embeddings (obtained by BERT) are used as the input features of bag-of-words and sequential models respectively. Using BERT helped us obtain better context-aware word embedding features so that we could improve the classification accuracy. A published BERT pre-trained model ("bert-base-uncased"[2]) is utilized as the embedding layer of LSTM model. "Bert-base-uncased" is a pretrained model on English language using a masked language modeling objective and its vocabulary size is 30,522. We set the maximum length of documents to 10,000 in the LSTM model because the average length of all the documents in the labeled dataset is about 10,000.

Since the text features and statistics feature are not compatible with each other, we will train models on these two sets of features separately. But we also try combining the results of these two sets of models to further boost up the prediction perfor-

[2]https://huggingface.co/bert-base-uncased

Model	Train Setting	Test Accuracy (Text)	Test Accuracy (Text + Statistics)
LR	300 (L)	57.58% (57/99)	58.59% (58/99)
RF	300 (L)	51.52% (51/99)	52.53% (52/99)
SVM	300 (L)	58.59% (58/99)	60.61% (60/99)
MLP	300 (L)	59.60% (59/99)	60.61% (60/99)
LSTM	300 (L)	61.62% (61/99)	63.64% (63/99)
LSTM	300 (L) + 2,170 (U)	61.62% (61/99)	63.64% (63/99)
DIVIDEMIX	300 (L) + 2,170 (U)	62.63% (62/99)	63.64% (63/99)
VI	300 (L) + 2,170 (U)	**66.67% (66/99)**	**67.68% (67/99)**
PL	300 (L) + 2,170 (U)	**71.72% (71/99)**	**75.76% (75/99)**

Table 4: Comparison on Train setting, Test Accuracy (Text), and Test Accuracy (Text + Statistics) between different eight trained models. VI is our variational inference based method, and PL is our peer loss based approach. 300 (L) means that 300 labelled dataset are used to train. 300 (L) + 2,170 (U) means that 300 labelled and 2,170 dataset are used to train.

Model	Precision	Recall	F1
LR	61.90%	50.98%	55.91%
RF	54.05%	39.22%	45.45%
SVM	63.04%	56.86%	59.79%
MLP	65.00%	50.98%	57.14%
LSTM	70.27%	50.98%	59.09%
DIVIDEMIX	65.11%	54.90%	59.57%
VI	**72.50%**	**56.86%**	**63.74%**
PL	**71.43%**	**88.24%**	**78.95%**

Table 5: Comparison on Precision, Recall, and F1 between different approaches (Setting: Text + Statistics)

mance. [3] A summation of their prediction probabilities will be used.

6.2 Results

The results of text only and text + statistics are reported in Table 4. From this table, we first observe that the ensemble models (combining text and statistics) outperform the ones trained only on text features. This suggests that the statistics feature are complementary to text feature.

We report that LR, RF, and SVM models (non-deep learning) trained using only statistics features are only able to obtain a 54.55%, 50.51%, and 56.57% test accuracy respectively. Therefore our experiments confirm that the performance of model training on text features is better.

We compare eight methods LR, RF, SVM, MLP, LSTM, DIVIDEMIX (Li et al., 2020), VI (our variational inference based method), and PL (our peer loss based method). The first five models are commonly used binary classification algorithms and they are trained only on 300 annotated data instances. VI and PL return the best performance and the result shows that our proposed methods consistently outperform other models. Among our two proposed approaches, PL obtains better performance and it reaches 75.76% accuracy. This is evidence to us that the PL approach works better in handling the noise; on the other hand, likely additional errors were introduced to VI during the process of estimating the error rates.

We also trained LSTM on both labeled and unlabeled datasets but with artificially provided labels. We observe the same performance as training only on the labeled dataset. It shows that the prediction performance cannot be improved if we do not use a noise-resistant procedure to correct the biases in the artificially provided labels.

The experimental results on Precision, Recall, and F1 score for eight models are also reported in Table 5. Our weakly supervised methods achieved the best performances consistently across different measures.

6.3 Ablation Study on Feature Importance for Research Replication

We explore which features are more indicative of an article's reproducibility. We perform the with/without experiments to compare the performance in different settings so that it can help us understand which features are more important in predicting replication.

The papers in our dataset contain different sections including title, authors, abstract, introduction, method, experiment, discussion, conclusion, ref-

[3] In the combination, the model using only statistics features is fixed to SVM since it has the best performance.

Model	Whole text	w/o Abs + Intro	w/o Method + Experiment	w/o Dis + Con + Ref + App
LR	**57.58% (57/99)**	54.55% (54/99)	51.52% (51/99)	57.58% (57/99)
RF	**51.52% (51/99)**	45.45% (45/99)	48.48% (48/99)	51.52% (51/99)
SVM	**58.59% (58/99)**	52.53% (52/99)	48.48% (48/99)	51.52% (51/99)
MLP	**59.60% (59/99)**	54.55% (54/99)	48.48% (48/99)	58.59% (58/99)
LSTM	**61.62% (61/99)**	58.59% (58/99)	42.42% (42/99)	60.61% (60/99)

Table 6: Accuracy comparison between different features on the test dataset

Donors **tend** to **avoid** charities that dedicate a high **percentage** of expenses to administrative and fundraising costs, limiting the ability of nonprofits to be effective. We propose a solution to this problem: Use donations from **major** philanthropists to cover overhead expenses and offer **potential** donors an overhead-free donation opportunity. A laboratory experiment testing this solution confirms that donations **decrease** when overhead increases, but only when donors **pay** for overhead themselves. In a field **experiment** with 40,000 potential donors, we compared the overhead-free solution with other **common** uses of initial donations. Consistent with **prior** research, informing donors that seed money has already been raised **increases** donations, as does a $1:$1 matching campaign. Our main result, however, clearly **shows** that informing **potential** donors that overhead **costs** 3 are covered by an initial donation significantly **increases** the donation **rate** by 80% (or 94%) and **total** donations by 75% (or 89%) compared with the seed (or matching) approach.

Table 7: Red color highlights words having positive weights and the absolute value is larger than 0.1. Blue color highlights words having negative weights and the absolute value is larger than 0.1. Classification result of Logistic Regression for this paper is Non-replicable (Wrong)

Donors tend to avoid **charities** that dedicate a **high** percentage of **expenses** to administrative and **fundraising** costs, **limiting** the ability of nonprofits to be effective. We **propose** a solution to this problem: Use **donations** from **major** philanthropists to **cover** overhead expenses and **offer potential donors** an overhead-free **donation** opportunity. A laboratory experiment testing this **solution confirms** that **donations decrease** when **overhead** increases, but only when **donors pay** for **overhead** themselves. In a field experiment with 40,000 potential donors, we compared the overhead-free solution with other **common** uses of initial donations. **Consistent** with **prior** research, **informing donors** that **seed money** has already been raised increases donations, as does a $1:$1 **matching** campaign. Our main result, however, clearly shows that **informing potential** donors that **overhead costs** 3 are covered by an initial donation **significantly** increases the donation rate by 80% (or 94%) and **total** donations by 75% (or 89%) compared with the **seed** (or matching) approach.

Table 8: Red color highlights words having positive weights and the absolute value is larger than 0.15. Blue colors highlight words having negative weights and the absolute value is larger than 0.15. Classification result of Peer Loss for this paper is Replicable (Correct)

erence, and appendix. We consider each section as a meta feature. The first set of features is title + authors + abstract + introduction, comprising the summary of this paper. The second set of features is methods + experiments which describe the details of the methods utilized in the paper and the effectiveness of the methods. The third set of features is discussion + conclusion + reference + appendix which consist the general conclusion and supplementary materials of this paper.

Experiments' results are reported in Table 6. We make several observations:

- Training using the entire body of text returns the best performance. This implies the necessity/informativeness of each component of an article.
- Removing the abstract and introduction leads to decreased performance but the reduction is not significant. Our conjecture is that the first set of features contains the summary of the whole paper, but it lacks details of methods and experiments.
- Cutting off the ending set of features (discussion+conclusion+reference+appendix) results in almost the same performance as the all text setting. This is primarily because the information in the third set of features has already been covered in the first set of features or is supplementary.
- Removing method+experiment leads to a significant reduction of testing accuracy. We conjecture this is because the second set of features contains the core details.

In summary, we found that the methods and experiments sections are more important than other sections.

6.4 Case Study

We showed two samples which have the same text but have different classification results with two different classifiers. The paragraph is selected from

the research paper "Avoiding overhead aversion in charity" published in Behavioral Economics. This article has been verified to be replicable. The goal of this case study is to provide an intuitive view about how the classifiers work and their ability to identify relevant contexts.

The classification result of LR classifier is non-replicable which is wrong. Since our text features are Tf-idf, there is a weight coefficient for each word in LR classifier. We highlight the words with larger weights in Table 7. As for the PL classifiers, its classification result is Replicable (Correct). We highlight the words with larger weights in Table 8. Because PL uses a neural network to train the model, there is a corresponding node in the input layer for each word. Each node has multiple links to the hidden layer and every link has a weight coefficient. For each code, we calculate the summation of all the weights. We do observe evidence that the PL classifier is able to capture more relevant keywords such as charity, donors, overhead, significantly, etc. This study demonstrates the possibility of using our works to identify the keywords or key paragraphs to spot-check an article.

7 Discussion

In this paper, we used two fields of corpus ("economic review" and "psychological science") to train our model together because both of them are social sciences that rely heavily on quantitative methodologies (e.g., survey, experiments) and draw conclusions based on statistics. Thus, they share the same definition of replicability such that whether the same statistical findings (e.g., effect size, p-value) can be reproduced in replications following the same methodological procedure with different samples. The same methodologies are also widely used in empirical sciences (e.g., lab experiments in Biology and Medicine) which demand replicability in the same sense and also follow the same format in reporting their procedures and findings. Thus, our proposed methods should also work in the contexts mentioned above.

8 Conclusion

The paper studies the possibilities of using weakly supervised learning methods based on text information of research papers to improve the prediction accuracy of research replication using a small amount of labeled data and a large amount of unlabeled data. Our experiments show that our approaches successfully improved prediction performance compared to the supervised models utilizing only statistic features and a small size of labeled dataset. Our approach can also be generically extended to other weakly supervised NLP.

Our study has limitations. First of all, our sampling of the unsupervised articles is not ideal. As a next step, we will include a more diverse and bigger pool of representative articles into our study. Our method replied on BERT for feature extraction, which remains largely as a "blackbox" processor. In the future, we plan to explore other advanced NLP techniques such as Named Entity Recognition, Relation Extraction, etc. to help us identify more explainable features. This information will help facilitate the human evaluation of a research claim's replicability.

Acknowledgments

This research is based upon work supported in part by National Science Foundation (NSF) under Grant No. RI-2007951 and the Defense Advanced Research Projects Agency (DARPA) and Space and Naval Warfare Systems Center Pacific (SSC Pacific) under Contract No. N66001-19-C-4014. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of NSF, DARPA, SSC Pacific or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

The authors would like to thank the members of the ReplicationMarkets team[4] for their helpful comments and suggestions. The list of members includes but is not limited to M. Bishop, Y. Chen, M. Gordon, T. Pfeiffer, R. Raab, C. Twardy, and J. Wang. The authors also would like to thank Dr. Bingjie Liu for her valuable feedback on the manuscript. We thank anonymous reviewers for valuable suggestions.

References

Adam Altmejd, Anna Dreber, Eskil Forsell, Juergen Huber, Taisuke Imai, Magnus Johannesson, Michael Kirchler, Gideon Nave, and Colin Camerer. 2019. Predicting the replicability of social science lab experiments. *PloS one*, 14(12).

[4] https://www.replicationmarkets.com/

Chris Biemann and Alexander Mehler. 2014. *Text mining: From ontology learning to automated text processing applications*. Springer.

Tiberiu Boroş, Stefan Daniel Dumitrescu, and Ruxandra Burtica. 2018. Nlp-cube: End-to-end raw text processing with neural networks. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 171–179.

Tom Bylander. 1994. Learning linear threshold functions in the presence of classification noise. In *Proceedings of the seventh annual conference on Computational learning theory*, pages 340–347.

Colin F Camerer, Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Johan Almenberg, Adam Altmejd, Taizan Chan, et al. 2016. Evaluating replicability of laboratory experiments in economics. *Science*, 351(6280):1433–1436.

Colin F Camerer, Anna Dreber, Felix Holzmeister, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Gideon Nave, Brian A Nosek, Thomas Pfeiffer, et al. 2018. Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nature Human Behaviour*, 2(9):637–644.

Nicolo Cesa-Bianchi, Shai Shalev-Shwartz, and Ohad Shamir. 2011. Online learning of noisy data. *IEEE Transactions on Information Theory*, 57(12):7907–7931.

Chih-Chung Chang and Chih-Jen Lin. 2011. Libsvm: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):27.

Open Science Collaboration. 2012. An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science*, 7(6):657–660.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

A Dreber, T Pfeiffer, E Forsell, D Viganola, M Johannesson, Y Chen, B Wilson, BA Nosek, and J Almenberg. 2019. Predicting replication outcomes in the many labs 2 study. *Journal of Economic Psychology*.

Charles R Ebersole, Olivia E Atherton, Aimee L Belanger, Hayley M Skulborstad, Jill M Allen, Jonathan B Banks, Erica Baranski, Michael J Bernstein, Diane BV Bonfiglio, Leanne Boucher, et al. 2016. Many labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, 67:68–82.

Leonard P Freedman, Iain M Cockburn, and Timothy S Simcoe. 2015. The economics of reproducibility in preclinical research. *PLoS Biol*, 13(6):e1002165.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep learning*. MIT press.

Tin Kam Ho. 1995. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Dan Jurafsky and James H Martin. 2014. Speech and language processing. vol. 3.

Richard A Klein, Kate A Ratliff, Michelangelo Vianello, Reginald B Adams Jr, Štěpán Bahník, Michael J Bernstein, Konrad Bocian, Mark J Brandt, Beach Brooks, Claudia Chloe Brumbaugh, et al. 2014a. Investigating variation in replicability. *Social psychology*.

Richard A Klein, Kate A Ratliff, Michelangelo Vianello, Reginald B Adams Jr, Štěpán Bahník, Michael J Bernstein, Konrad Bocian, Mark J Brandt, Beach Brooks, Claudia Chloe Brumbaugh, et al. 2014b. Theory building through replication: Response to commentaries on the "many labs" replication project.

Richard A Klein, Michelangelo Vianello, Fred Hasselman, Byron G Adams, Reginald B Adams Jr, Sinan Alper, Mark Aveyard, Jordan R Axt, Mayowa T Babalola, Štěpán Bahník, et al. 2018. Many labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, 1(4):443–490.

Junnan Li, Richard Socher, and Steven CH Hoi. 2020. Dividemix: Learning with noisy labels as semi-supervised learning. *arXiv preprint arXiv:2002.07394*.

Qiang Liu, Jian Peng, and Alexander T Ihler. 2012. Variational inference for crowdsourcing. In *Advances in neural information processing systems*, pages 692–700.

Yang Liu and Hongyi Guo. 2020. Peer loss functions: Learning from noisy labels without knowing noise rates. *International Conference on Machine Learning*.

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. 2018. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993.

Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. 2013. Learning with noisy labels. In *Advances in neural information processing systems*, pages 1196–1204.

Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. 2018. Realistic evaluation of deep semi-supervised learning algorithms. In *Advances in Neural Information Processing Systems*, pages 3235–3246.

H Pashler, B Spellman, S Kang, and A Holcombe. 2019. Psychfiledrawer: archive of replication attempts in experimental psychology. *Online¡ http://psychfiledrawer. org/view_ article_list. php*.

Chao-Ying Joanne Peng, Kuk Lida Lee, and Gary M Ingersoll. 2002. An introduction to logistic regression analysis and reporting. *The journal of educational research*, 96(1):3–14.

Adam Russell. 2019. Systematizing confidence in open research and evidence (score). Technical report, Tech. Rep., Defense Advanced Research Projects Agency, Arlington, VA.

Clayton Scott. 2015. A rate of convergence for mixture proportion estimation, with application to learning from noisy labels. In *Artificial Intelligence and Statistics*, pages 838–846.

Clayton Scott, Gilles Blanchard, and Gregory Handy. 2013. Classification with asymmetric label noise: Consistency and maximal denoising. In *Conference On Learning Theory*, pages 489–511.

Yusuke Shinyama. 2014. Pdfminer.

Daniel J Simons, Alex O Holcombe, and Barbara A Spellman. 2014. An introduction to registered replication reports at perspectives on psychological science. *Perspectives on Psychological Science*, 9(5):552–555.

Uri Simonsohn. 2015. Small telescopes: Detectability and the evaluation of replication results. *Psychological science*, 26(5):559–569.

Brendan Van Rooyen, Aditya Menon, and Robert C Williamson. 2015. Learning with symmetric label noise: The importance of being unhinged. In *Advances in Neural Information Processing Systems*, pages 10–18.

Thomas Wolf, L Debut, V Sanh, J Chaumond, C Delangue, A Moi, P Cistac, T Rault, R Louf, M Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv, abs/1910.03771*.

Yang Yang. 2018. The replicability of scientific findings using human and machine intelligence. https://www.metascience2019.org/presentations/yang-yang/ Metascience 2019.

Zhi-Hua Zhou. 2018. A brief introduction to weakly supervised learning. *National Science Review*, 5(1):44–53.