Predicting the Reproducibility of Psychological Research:

A Text Classification Approach to Academic Paper Abstracts

Abhilasha A. Kumar[1]

[1]Washington University in St Louis, MO

Abstract

A recent replication of 100 psychological studies, conducted by the Open Science Collaboration (2015) revealed that the reproducibility of psychological research is far lower than desirable, and also suggested that variables such as the sub-discipline of psychology, the original strength of evidence, and the extent to which the original effect was surprising in fact predict the likelihood of replication for a particular study. This paper presents a different approach to the problem of predicting reproducibility, by analyzing the text in the abstracts of 91 of the original studies to investigate whether specific word use or textual characteristics in the abstracts of scientific papers can provide insight into the extent to which they will replicate. A latent semantic analysis of the abstracts, followed by text-classification revealed that the textual content in the abstracts can predict the reproducibility of a particular study with 70% accuracy, and that specific word use may be indicative of findings that are likely to replicate. Additionally, there was no evidence of lexical diversity or parts-of-speech usage predicting the likelihood of replication, although the longer length of the abstract did predict replication status for the abstract. These findings suggest that specific linguistic properties in the abstract text may not be particularly informative in and of itself in predicting replication, although the specific types of words used in the abstract might provide further insight into the type of study, and the likelihood of replication, as shown by the text classification approach.

Replicability is a critical, defining feature of science, and whether psychological science meets this criterion of reproducibility is an open question that has received considerable attention in the recent news. A reproducibility project conducted by the Open Science Collaboration (2015) attempted to replicate 100 psychological studies published in high-ranking psychology journals and reported that only 36% of the original results were statistically significant in the replication attempt. This project sparked off an important debate in the field of psychology regarding the value of statistical significance (Wasserstein & Lazar, 2016), the publication bias (Bialystok, Kroll, Green, MacWhinney, & Craik, 2015; Francis, 2012), and the need for newer, more powerful statistical tools to assess reproducibility in psychological research (Etz & Vandekerckhove, 2016).

An important aspect of scientific research is accurately communicating scientific findings, but the influence of language in academic communication has not been adequately studied, even though there is evidence to show that specific word use and linguistic styles shape perception and inheritance of cultural norms (Gelman & Roberts, 2017). It is possible that academic papers reflect specific linguistic characteristics that may be an indicator of their likelihood to replicate. This paper presents a new approach to understanding the replication crisis by analyzing the text of 91 of the original studies in the reproducibility project, and exploring whether any linguistic variables predict whether a particular abstract will replicate in the future. We explore this question through latent semantic analysis and text classification, and also look at the impact of lexical diversity and parts-of-speech usage on replication.

Method

Materials

Plain texts abstracts of 91 of the original studies from the reproducibility project (Open Science Collaboration, 2015) were extracted and used as the primary corpus for analyses in this study. Information about whether a study was replicated or not was obtained from an openly available dataset shared by the Center of Open Science at https://osf.io/ezcuj/files/. This dataset provides information about authors and co-authors of the original studies, results of the replication attempt, "surprisingness" of the result and several other variables relevant to the original studies in the reproducibility project.

Procedure

Importing and Sorting Texts. All abstract texts were imported into Python 3.4 and merged with replication data from the Center for Open Science to form a collection of abstract texts and a label that indicated whether the corresponding study replicated (1) or failed to replicate (0). This composite dataset was then used in all further analyses. Figure 1 displays the code used to import and sort the texts.

Calculating Linguistic Properties. Lexical diversity was calculated for each text, before and after removing punctuations and stopwords from the English language. Numbers were not removed from texts at this point since authors may frequently cite other authors, which may improve lexical diversity and may contribute to the likelihood of replication. Each token in the text was also tagged using a parts-of-speech (POS) tagger available through the *nltk* package in Python. Figure 2 displays the code used to calculate lexical diversity and perform POS tagging on the texts.

Splitting the Corpus into Training and Test Set. The complete corpus of 91 texts was randomly split into a training and test set based on a 70-30 percent ratio. The classifier was first trained using the abstracts and replication labels in the training set (68 texts), and was then applied to the remaining texts in the corpus (23 texts) to evaluate classification accuracy.

Applying Latent Semantic Analysis. Before proceeding to the classification stage, all texts were processed and converted into term-document matrices using latent semantic analysis. This process involved three stages: (1) converting the raw text into a Term-Frequency (*tf*) matrix, which results in a matrix that tabulates the frequency of every term in a single text; (2) calculating the Inverse Document Frequency (*idf*), which is the logarithm of number of texts in the total corpus divided by the number of texts in which the particular term appears; (3) Multipliying the *tf* and *idf* matrices to ensure that each term is appropriately weighted by its relative frequency of occurrence and importance in the corpus; (4) Applying truncated singular value decomposition to the resulting tfidf matrix decompose the matrix into meaningful components and simultaneously reduce the dimensionality of the matrix. The final result from these analyses is a set of components derived from the terms in the corpus, and each text in the corpus expressed as a linear combination of these components. Figure 3a shows the five components used in this analyses as a linear combination of the terms in the training set, and Figure 3b shows each text in the corpus as a linear combination of the five components. After applying latent semantic analysis to the training set, the test set was transformed to fit the latent semantic space generated by the training set.

Classifying Text. Following the singular value decomposition of the training and test sets, the resulting matrices were passed to a batch classifier that used 8 different classifiers to classify the texts in the test set as replicable or not replicable. The classifiers used in this study

were logistic regression, k-nearest neighbors, linear support vector machine (SVM), gradient boosting classifier, decision tree classifier, random forest classifier, neural net classifier and Naïve Bayes classifier. The code used to implement these classifiers is available from http://ataspinar.com/2017/05/26/classification-with-scikit-learn/. The final results from these analyses included the overall performance of each classifier, on the training and test set, total time taken to perform classification and the final result of classification for each text in the test set.

Results

Lexical Diversity. Figure 4 displays the lexical diversity of all texts in the corpus as a function of their replication status. A generalized linear regression model with probability of replication as a binomial outcome variable, and lexical diversity as a predictor revealed that lexical diversity does not predict the likelihood of replication, odds ratio = 0.19, $p = 0.62$. Interestingly, when the length of the abstract is added as an additional predictor to the model, lexical diversity does not significantly predict the likelihood of replication, suggesting that the richness of the abstract is not a potentially strong indicator of replication. However, length (in number of words) predicts likelihood of replication, odds ratio = 1.02, $p = .042$. These results suggest that longer abstracts have a greater likelihood of replication. Figure 5 displays the predicted probabilities of replication as a function of abstract length, as predicted by the regression model.

Parts of Speech Usage. Figure 6 displays the percent occurrence of adjectives, nouns, verbs and other parts of speech in all texts of the corpus as a function of their replication status. A generalized mixed effects regression model with probability of replication as a binomial outcome variable and percentage occurrence, and type of POS as predictors revealed that percent

occurrence of parts of speech in the abstracts does not predict the likelihood of replication, $\chi 2_{(3)} = .053$, $p = .997$.

Classifier Performance. Table 1 displays the classification accuracy and completion times for all classifiers, based on one sample of the training and test set. Most classifiers were at-chance prediction, suggesting that the likelihood of replication was difficult to predict from the text in the abstracts, but the logistic regression, linear SVM and neural net classifiers performed the best with approximately 70% accuracy. Figure 7 displays the performance of the k-nearest neighbor classifier based on one run of the corpus, since it was not a good classifier and performed below-chance on these data. It can be seen that this classifier was generally good at predicting when an abstract will not replicate, but did not perform well when predicting replication for successfully replicated abstracts. Using the package eli5, the contribution of specific features from a document in the replication process was also examined for the logistic regression classifier. Figure 8 displays results from 4 abstract texts when the classifier made the correct decision, with the probability of replication and the specific features that contributed positively (in green) and negatively (in red) to the classifier decision. Interestingly, words such as “experimental”, “successful”, “empirical” and “evidence” seem to indicate higher probability of replication, whereas phrases such as “may”, “possible”, “may indirectly lead to” in the abstract seem to indicate lower probability of replication by the classifier. These results suggest that using language that signifies experimental work and concrete evidence seems to be an indicator of successful future replication, whereas hedging claims with qualifiers like “may” and “possible” likely indicates that the study will not replicate.

Conclusion

The current study investigated the influence of specific linguistic features on the likelihood of replication, using abstract text from 91 psychology papers published in high-ranking scientific journals. We found no evidence of lexical diversity or parts of speech usage in predicting the likelihood of replication. The length of the abstract emerged as a significant predictor of replication, suggesting that abstracts of papers that explain their work in detail are possibly more likely to replicate. However, this effect was small and so further work would be needed to clarify the effect of length on the likelihood of replication. Further, based on classifier performance, it appears that specific words such as “experimental” and “successful” might be weak, but important signatures of replication, although several of the classifiers were also at-chance performance, suggesting that abstract text in and of itself may not be a very powerful predictor of future replication.

References

Bialystok, E., Kroll, J. F., Green, D. W., MacWhinney, B., & Craik, F. I. M. (2015). Publication Bias and the Validity of Evidence: What's the Connection? *Psychological Science*, *26*(6), 944–946. https://doi.org/10.1177/0956797615573759

Gelman, A., & Carlin, J. (2017). Some Natural Solutions to the p-Value Communication Problem—and Why They Won't Work. *Journal of the American Statistical Association*, *112*(519), 899–901. https://doi.org/10.1080/01621459.2017.1311263

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's Statement on *p* -Values: Context, Process, and Purpose. *The American Statistician*, *70*(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108

Gelman, S. A., & Roberts, S. O. (2017). How language shapes the cultural inheritance of categories. *Proceedings of the National Academy of Sciences*, *114*(30), 7900–7907. https://doi.org/10.1073/pnas.1621073114

Etz, A., & Vandekerckhove, J. (2016). A Bayesian perspective on the reproducibility project: Psychology. *PLoS ONE*, *11*(2), 1–12. https://doi.org/10.1371/journal.pone.0149794

Francis, G. (2012). Publication bias and the failure of replication in experimental psychology. *Psychonomic Bulletin and Review*, *19*(6), 975–991. https://doi.org/10.3758/s13423-012-0322-y

Aarts, A. A., Anderson, J. E., Anderson, C. J., Attridge, P. R., Attwood, A., Axt, J., … Zuni, K. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251). https://doi.org/10.1126/science.aac4716

Tables and Figures

Table 1

Classifier performance accuracy for abstract texts

Classifier	Training Score	Test Score	Train Time (s)
Logistic Regression	0.588	0.696	0.006
Linear SVM	0.574	0.696	0.002
Neural Net	0.588	0.696	0.095
Gradient Boosting Classifier	1.000	0.522	0.308
Decision Tree	1.000	0.522	0.001
Random Forest	1.000	0.478	1.139
Naive Bayes	0.662	0.478	0.001
Nearest Neighbors	0.765	0.304	0.002

```
from sklearn.model_selection import train_test_split

# dictionary of text files: encodings and labels (e.g. 0 for not replicated, 1 for replicated)
textfiles = {

  "1.txt": {"encoding": "ISO-8859-1", "label":1},
"3.txt": {"encoding": "ISO-8859-1", "label":0},
"99.txt": {"encoding": "ISO-8859-1", "label":0},
"100.txt": {"encoding": "ISO-8859-1", "label":0},
"101.txt": {"encoding": "ISO-8859-1", "label":0},
"102.txt": {"encoding": "ISO-8859-1", "label":0},

}

# read in each file in the appropriate encoding using utf-8 as the default encoding
texts = []
labels = []
for textfilename, attributes in textfiles.items():
    f = open(textfilename, 'r', encoding=attributes.get('encoding', 'utf-8'))
    texts.append(f.read())
    labels.append(attributes['label'])

# test/train split the feature and label data randomly into train and test data
X_train, X_test, y_train, y_test = train_test_split(texts, labels)
```

Figure 1. Python code demonstrating the procedure used to import abstract text and replication status and split into training and test set for classification.

```
def lexical_diversity(abstract):
    return len(set(abstract))/len(abstract)
```

```
from nltk import word_tokenize
from nltk.corpus import stopwords
import pandas as pd
punctuations = ['(',')',';',':','[',']',',', '.']
stop_words = stopwords.words('english')
pd_ld = pd.DataFrame({'Text Number': [], 'Lexical Diversity': [], 'Text': [], 'Length': []})
pd.set_option('display.max_colwidth', -1)

for i in range(0, len(texts)):
    tokens = word_tokenize(texts[i])
    keywords = [word for word in tokens if not word in stop_words and not word in punctuations]
    ld = lexical_diversity(keywords)
    d = {'Text Number': [i+1], 'Lexical Diversity': [ld], 'Text': [keywords], 'Length': [len(keywords)]}
    df1 = pd.DataFrame(data=d)
    pd_ld = pd_ld.append(df1)

pd_ld
pd_ld.to_csv('lexical_diversity_python.csv', sep = ",")
```

```
pd_ld
```

	Length	Lexical Diversity	Text
0	163.0	0.515337	Researchers interested in counterfactual thinking have often found that upward counterfactual thoughts\nlead to increased motivation to improve in the future, although at the cost of increased negative affect.\nThe present studies suggest that because upward counterfactual thoughts indicate reasons for a poor\nperformance, they can also serve as excuses. In this case, upward counterfactual thoughts should result\nin more positive self-esteem and reduced future motivation. Five studies demonstrated these effects in the\ncontext of self-handicapping. First, upward counterfactual thinking was increased in the presence of a\nself-handicap\nprotected self-esteem following failure. Finally, upward counterfactual thoughts that protect self-esteem\nreduced preparation for a subsequent performance as well as performance itself. These findings suggest\nthat the consequences of upward counterfactuals for affect and motivation are moderated by the goals of\nthe individual as well as the content of the thoughts.\n

```
### POS tagging
import nltk
from nltk import word_tokenize
from nltk.corpus import stopwords
import pandas as pd
punctuations = ['(',')',';',':','[',']',',', '.']
stop_words = stopwords.words('english')
pd_pos = pd.DataFrame({'Text Number': [], 'Adjective': [], 'Noun': [], 'Verb': [],
                       'Other': [], 'Text': [], 'Length': []})
pd.set_option('display.max_colwidth', -1)

for i in range(0, len(texts)):
    tokens = word_tokenize(texts[i])
    keywords = [word for word in tokens if not word in stop_words and not word in punctuations]
    pos_tagger = nltk.pos_tag(keywords)
    adj = 0
    noun = 0
    verb = 0
    other = 0
    for term, pos in pos_tagger:
        if pos == "JJ":
            adj = adj + 1
        elif pos == "NN" or pos == "NNS":
            noun = noun + 1
        elif pos == "VBP" or pos == "VB" or pos == "VBG":
            verb = verb + 1
        else:
            other = other +1

    d = {'Text Number': [i+1], 'Adjective': [adj], 'Noun': [noun], 'Verb': [verb],
         'Other': [other], 'Text': [keywords], 'Length': [len(keywords)]}
    df1 = pd.DataFrame(data=d)
    pd_pos = pd_pos.append(df1)

pd_pos
#pd_pos.to_csv('pos_python_code.csv', sep = ",")
```

Figure 2. Python code demonstrating the procedure used to calculate lexical diversity and count the number of nouns, adjectives, verbs and other POS in the abstract text.

	component_	component_	component_	component_	component_4
Although	0.24583974	0.12231557	-0.2449197	0.03015473	-0.0216192
Low-	0.30915294	-0.2186808	-0.1709363	0.4298985	-0.0206538
Rabbi	0.17349841	0.00362415	9.81E-05	-0.04862	-0.035098
The error-	0.20828157	-0.0936966	-0.058658	0.12679978	0.15076981
Recent	0.19095355	0.14021459	0.19488515	-0.0691712	-0.195426

Figure 3a. Each abstract as a linear combination of 5 components, obtained after latent semantic analysis on the abstract corpus.

	-PRON-	ab	ability	able	abovechance	absence
component_	0.18664624	0.00854588	0.04377457	0.02713061	0.00526389	0.00977676
component_	-0.0752596	0.00953583	-0.0110849	0.01460218	0.0132627	0.02187891
component_	0.04229962	0.00734071	-0.026813	0.02465795	0.01814378	0.00195442
component_	0.00919433	0.01567538	0.09189124	-0.0024599	0.01784382	-0.0152355
component_	-0.0491418	-0.017918	0.00292112	0.01029141	-0.011106	0.0133903

Figure 3b. Each component as a linear combination of the features extracted from the abstract corpus via latent semantic analysis.

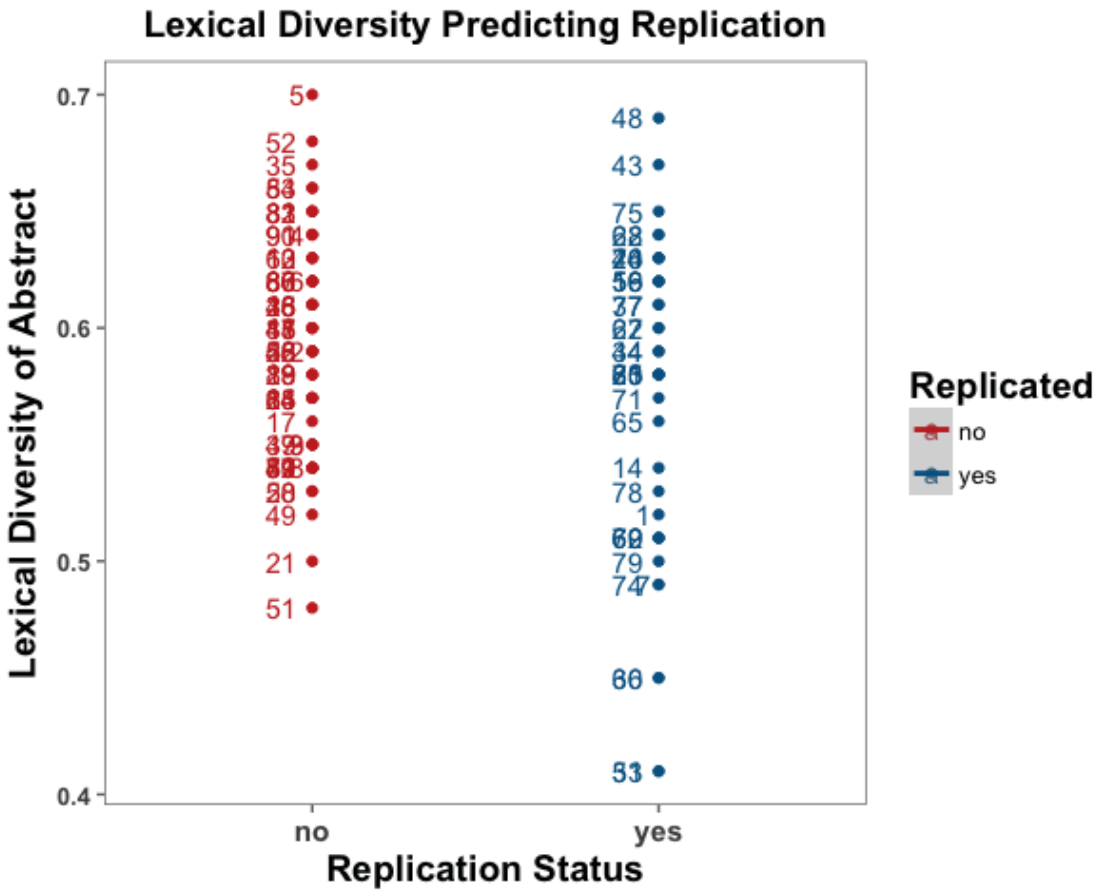

Figure 4. Replication status of the abstract as a function of lexical diversity of the abstract.

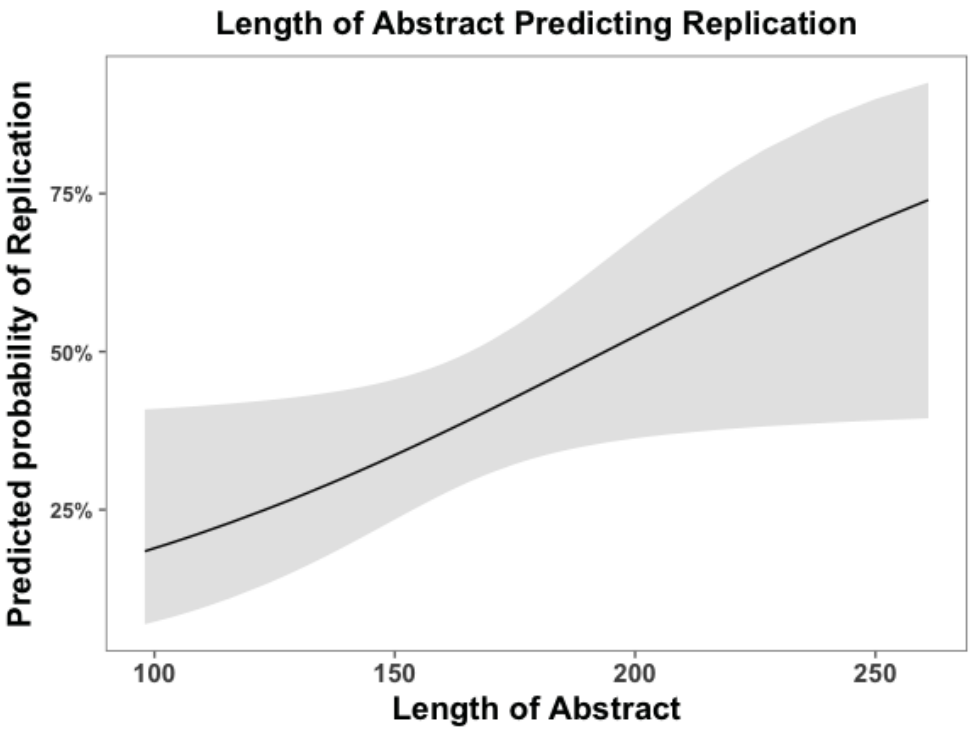

Figure 5. Replication status of the abstract as a function of the length of the abstract (in number of words), as predicted by the generalized linear logistic regression model.

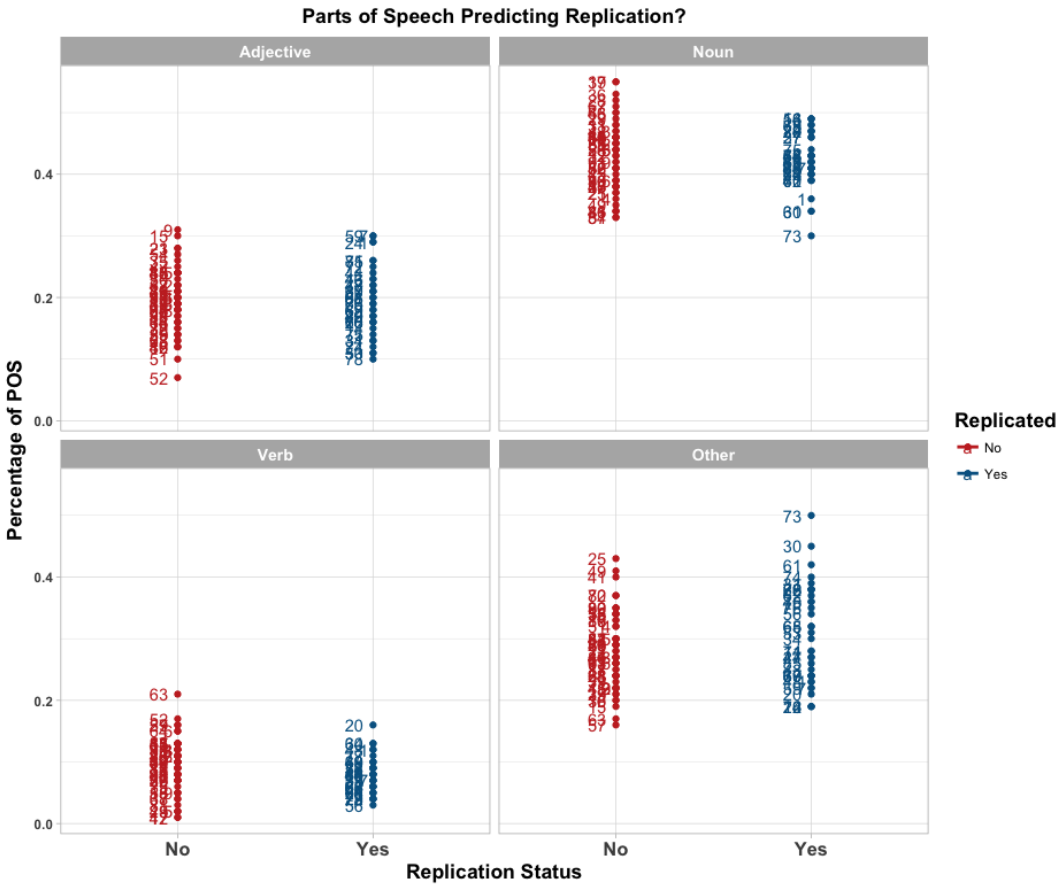

Figure 6. Replication status of the abstract as a function of the percentage usage of adjectives, nouns, verbs and other parts of speech in the text.

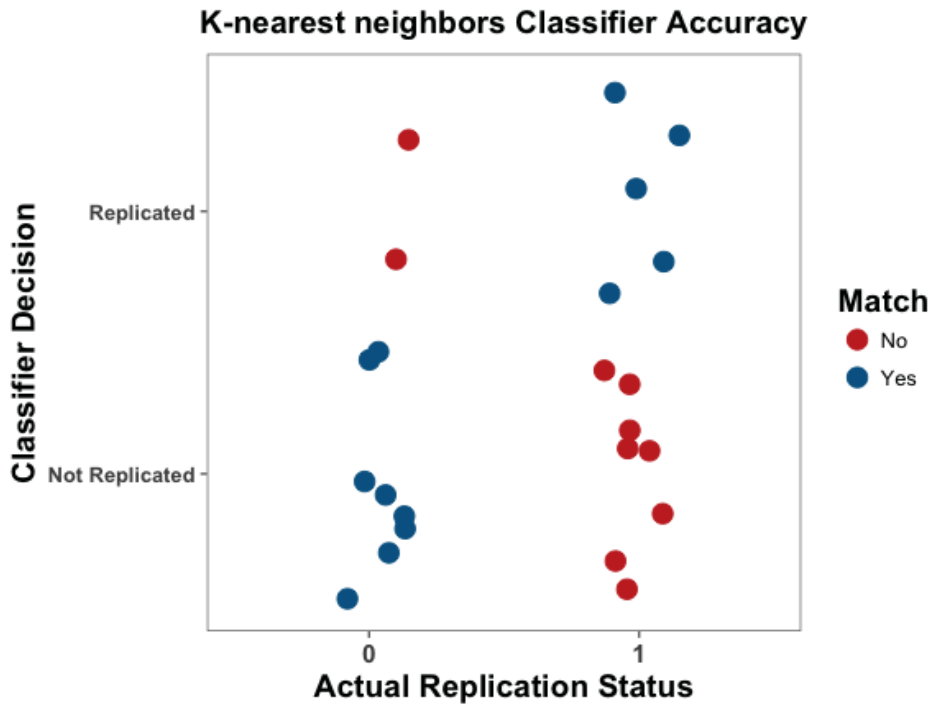

Figure 7. Performance of the k-nearest neighbors classifier as a function of actual replication status for the abstract texts. A "match" indicates when the classifier correctly identified an abstract as "replicated" or "not replicated".

y=replicated (probability **0.603**, score **0.417**) top features

Contribution?	Feature
+0.457	Highlighted in text (sum)
-0.041	<BIAS>

empirical evidence on selective exposure to information after decisions is contradictory: whereas many studies have found a preference for information that is consistent with one's prior decision, some have found a preference for inconsistent information. the authors propose that different available information quantities moderate these contradictory findings. four studies confirmed this expectation. when confronted with 10 pieces of information, decision makers systematically preferred decision-consistent information, whereas when confronted with only 2 pieces of information, they strongly preferred decision-inconsistent information (study 1). this effect was not due to differences in processing complexity (study 2) or dissonance processes (study 3) but could be traced back to different salient selection criteria: when confronted with 2 pieces of information, the salient selection criterion was information direction (consistent vs. inconsistent), which caused a preference for inconsistent information. in contrast, when confronted with more than 2 pieces of information, the salient selection criterion was expected information quality, which caused a preference for consistent information (study 4). (psycinfo database record (c) 2016 apa, all rights reserved)

y=replicated (probability **0.636**, score **0.559**) top features

Contribution?	Feature
+0.747	Highlighted in text (sum)
-0.188	<BIAS>

previous research on the communication of emotions has suggested that bargainers obtain higher outcomes if they communicate anger than if they communicate happiness because anger signals higher limits, which in turn leads opponents to give in. building on a social functional account of communicated emotions, the authors demonstrate that the behavioral consequences of communicated anger strongly depend on structural characteristics of the bargaining situation. the results of 3 experimental studies on ultimatum bargaining corroborate the notion that communicated anger signals higher limits and that emotion effects are contingent on bargainersõ expectation that low offers will be rejected. the data also indicate, however, that communicating anger in bargaining may backfire. the findings suggest that bargainers who communicate anger may obtain lower outcomes (a) when their opponent has a possibility to deceive them during bargaining and (b) when the consequences of rejecting their opponentõs offer are low. taken together, the current article reveals the boundary conditions of successful communication of anger in bargaining.

y=not replicated (probability **0.645**, score **-0.596**) top features

Contribution?	Feature
+0.560	Highlighted in text (sum)
+0.036	<BIAS>

people are motivated to maintain social connection with others, and those who lack social connection with other humans may try to compensate by creating a sense of human connection with nonhuman agents. this may occur in at least two waysñby anthropomorphizing nonhuman agents such as nonhuman animals and gadgets to make them appear more humanlike and by increasing belief in commonly anthropomorphized religious agents (such as god). three studies support these hypotheses both among individuals who are chronically lonely (study 1) and among those who are induced to feel lonely (studies 2 and 3). additional findings suggest that such results are not simply produced by any negative affective state (study 3). these results have important implications not only for understanding when people are likely to treat nonhuman agents as humanlike (anthropomorphism), but also for understanding when people treat human agents as non-human (dehumanization).

y=not replicated (probability **0.606**, score **-0.431**) top features

Contribution?	Feature
+0.313	Highlighted in text (sum)
+0.117	<BIAS>

it has been claimed that bilingualism enhances inhibitory control, but the available evidence is equivocal. the authors evaluated several possible versions of the inhibition hypothesis by comparing monolinguals and bilinguals with regard to stop signal performance, inhibition of return, and the attentional blink. these three phenomena, it can be argued, tap into different aspects of inhibition. monolinguals and bilinguals did not differ in stop signal reaction time and thus were comparable in terms of active-inhibitory efficiency. however, bilinguals showed no facilitation from spatial cues, showed a strong inhibition of return effect, and exhibited a more pronounced attentional blink. these results suggest that bilinguals do not differ from monolinguals in terms of active inhibition but have acquired a better ability to maintain action goals and to use them to bias goal-related information. under some circumstances, this ability may indirectly lead to more pronounced reactive inhibition of irrelevant information.

Figure 8. Specific prediction probabilities and top features for 4 abstracts on which the logistic regression classifier correctly predicted the replication status.