Detection of Fraudulent Science Using Linguistic Cues

Jiri Munich†, Bennett Kleinberg‡, Bruno Verschuere†, David Markowitz††, and Jeffrey Hancock‡‡
†Department of Psychology
University of Amsterdam
j.munich@uva.nl
‡Department of Security and Crime Science
University College London
††School of Journalism and Communication
University of Oregon
‡‡Department of Communication
Stanford University

Abstract

We use texts of scientific articles to train an algorithm for the detection of scientific fraud. Kernel Trick Support Vector Machine shows 60% classification accuracy, Random Forests perform at 80% classification accuracy. In an exploratory analysis, we try to identify the most important features. Most of the identification accuracy seems attributable to named entity counts.

1 Introduction

Science is a large and complex system producing an ever increasing number of publications (Bornmann and Mutz, 2015). The availability of large research-related data sets allowed the emergence of *Science of Science* (Fortunato et al., 2018), a systematic study of the mechanisms and problems involved in the production of knowledge. Some recent empirical work has raised concerns about the reliability of many published findings. For instance, many results published in top psychology journals could not be replicated (Open Science Collaboration, 2015) and only one third of basic research in drug development passes phase II clinical trials (Thomas et al., 2016). Many factors might be contributing to this lack of reproducibility, among others questionable research practices or outright fraud.

1.1 Questionable research practices and research fraud

The common assumption that science is self-correcting has been put under question. The incentives in science encourage many problematic practices prioritizing the attractiveness of research findings over their verisimilitude (Nosek et al., 2012). Funding is strongly linked to publishing output, which in turn depends on having interesting and positive finding, forming *perverse incentives* for dubious academic behaviour (Angell, 1986). Ioannidis (2012) makes a distinction between questionable research practices (QRPs) and outright fraud. Admittedly, the line separating QRPs and fraud can be blurry and field-dependent. QRPs refer to arbitrary choices motivated only by achieving the desired outcome (optional stopping of data collection when statistical test returns significant results, arbitrary exclusion of observations, p-hacking... (John et al., 2012)) that can still be deemed acceptable in the field, while data fraud refers to a direct manipulation or fabrication of data and/or results. Most QRPs can be identified by the examination of methods and analysis sections of articles. Fraud, on the other hand, is harder to spot and fraudulent articles then both contaminate the literature with false findings and introduced uncertainty about genuine publications, as a reader might not be able to distinguish one from the other.

Estimated 2% of scientists admit to committing fraud themselves and 14% have knowledge of fraud committed by their colleagues (Fanelli, 2009). The total number of retractions for fraud has been increasing over time (Steen, 2011) and the proportion of articles retracted for fraud on all published literature has increased almost 10 times between years 1975 and 2011 to about .01% of all published literature with scientific fraud accounting for about 43% of all retractions in that period (Fang et al., 2012).

The aim of this article is to develop an algorithm to detect fraudulent scientific publications using linguistic cues.

1.2 Detecting scientific fraud

Sometimes, data fraud gets discovered when it is reported by a knowledgeable informant (e.g. Jiao et al. (2013)). However, there have been cases of fraud discovered by previously uninformed readers. Misconduct by dr. Jens Forster at the Univer-

sity of Amsterdam was first brought to attention by a whistleblower analysis of descriptive statistics form three Forster's papers (Whistleblower Report, 2012). A follow-up investigation of Forster's publications has concluded that the consistency of reported statistics is so high it is unlikely to have come from genuine data (Peeters et al., 2015). A statistical method developed by Simonsohn (2013) uses reported means and standard deviations, or in better cases raw data, to highlight suspiciously unlikely patterns. Other methods based on identifying errors in reported statistics are so easily implementable they are available online to test any scientific publication (Epskamp et al.).

Detection tools based on reported statistics can be powerful, but they have several limitations. Chambers (2017) lists twelve ways how a fraudulent scientist can avoid the detection of his misconduct. For instance, an experimenter can switch observations between conditions to achieve the desired results, while maintaining natural variation within his data. Even more simply, when the raw data are no longer accessible, it becomes harder to confirm any suspicion based on reported descriptive statistics.

1.3 The language of deception

Unlike original data that often remain undisclosed (Houtkoop et al., 2018) every scientific publication is formulated in written language. The language used for reporting fraudulent findings is therefore available for analysis (as long as the researcher can secure access to the pay walled articles).

Computational linguistics have been successfully used for detecting deception with 75% accuracy in online communication (Ho et al., 2016; Ho and Hancock, 2018), 80% in reported intentions (Kleinberg et al., 2018) or 74% accuracy in fake news detection (Pérez-Rosas et al., 2017). All of these results are far above human deception detection accuracy, which was estimated to average 54% across contexts (Bond Jr and DePaulo, 2006) or at 52% specifically for deception in written reports (Masip et al., 2012).

Markowitz and Hancock (2014) have applied computational linguistics on deception in scientific reports. Focusing on the case of a Dutch psychologist Diedrik Stapel who was implicated with repeated data fraud, the authors used the results of an institutional investigation (Levelt Committee et al., november 2012) to compare the language of 24 fraudulent and 25 genuine publications. The presence of various discourse markers such as indicators of certainty (e.g. words "explicit", "precise) or certain adjectives (e.g. "cooperative") could predict fraud at 71% accuracy. Moving from a single author cases study, Markowitz and Hancock (2016) processed papers from the PubMed database retracted for fraud between years 1973-2013. All fraudulent publications were matched with similar non-retracted publications and compared on their *obfuscation index* scores a variable indicating the use of scientific jargon, abstraction and other features that might be used in deceptive language. Fraudulent papers displayed significantly higher use of obfuscating discourse across article sections. However, when using the obfuscation features as predictors of fraud, the accuracy was close to chance level at 57%.

1.4 The present study

Our research question is, whether linguistic features betray scientific fraud. Contrary to deductive approaches commonly used in psychology, we focus on prediction rather than explanation (Yarkoni and Westfall, 2017). We use machine-learning approach mainly because it allows the testing of theories that, due to the subject matter complexity, can make predictions only on a vaguer level (e.g. "fraudulent authors will combine modal adverbs differently than other authors"), but not more specific (e.g. "there is a particular pattern of specific modal adverbs used only by fraudulent authors").

The use of k-fold cross-validation and hold-out sets also prevents some problems related to overfitting. Given these advantages, it appears that features used for prediction could be selected without much theorizing. However, features motivated by theory might not only increase the predictive accuracy of an algorithm, but also broaden the scope of feature selection. In the following sections, we review both exploratory features and more reasoned concept-related features that could betray academic fraud.

1.4.1 Exploratory features

N-grams. Some information about the writing style of an author can be contained in the word use, both vocabulary and word combinations. Word use can be capture by n-grams. N-grams are n-long sequences of words occurring in a text. For instance a sentence "Results

are valid." could be decomposed into three unigrams ("results", "are", "reliable"), two bigrams ("results_are", "are_reliable") or one trigram ("results_are_reliable").

N-grams with stopwords. To capture the semantic dimensions of a text, n-grams can be computed without stopwords, that is, words that have no meaning on their own (e.g.*"a"*,*"the"*), as stopwords conceal some co-occurrences. On the other hand, stopwords can indicate meaningful information about syntax (for instance for an observation vs. for the observation). Overlapping features between the two sets of n-grams were removed from the stopwords free n-grams. Therefore, we use two sets of n-grams: with and without stopwords.

1.4.2 Reasoned concept-related features

We argue, that authors reporting fraud can either intentionally or unintentionally follow specific writing strategies that can ultimately betray their fraud. They could be attempting linguistic obfuscation decreasing the *readability* of their reports, *hedging* or a using specific informational structures indicated by their *syntax*. Besides these, there might be predictors relevant even to cases, when the writing author is not actually aware of the fraud (for instance when fraud is committed by a co-author).

Readability. Fraudulent scientists can try to confuse or impress the reader using obfuscating language with complex sentences and expert terminology (Markowitz and Hancock, 2016). Readability can be quantified, for instance with the Flesch index based on a) sentence length; b) word length; (3) personal word count (e.g. nouns with natural gender); (4) personal sentence count (Flesch, 1948)

Hedging. Another available strategy is *hedging*, where a certain level of uncertainty is indicated by the authors in the anticipation of possible criticism (Hyland, 1998). An author who writes *"It seems that our data support the conclusion"* signals a different level of confidence and willingness to discuss the findings than an author who writes *"Our data support the conclusion"*. People use specific types of words to indicate their epistemic stance (Tseronis et al., 2009; Gablasova et al., 2017) and fraudulent scientists might either express too much confidence or admit too much subjectivity about their data.

Syntax. The way sentences are built might reflect the way an author conveys information. A sentence can consist of a topic and a comment. The topic indicates what is being talked about and comment what is being said about the topic (Jacobs, 2001). Automated topic-comment is difficult to implement. However, English language displays a major correlation between the topic of a sentence and its subject (Lambrecht, 1996). Pairs such as subject-object relate to purely grammatical categories, while topic-comment speak about the meaning of the sentence. As such, topic-comment of a sentence can differ from the subject-object, but the subject-object pair can be used as its proxy. We argue that the grammatical structure of sentences in a text can be used as an indicator of strategies used to convey information used by the author. We focus on the object, subject and the root of a sentence.

To clarify, we can look at the meanings of root, subject and object. At the core of the sentence is the root a verb indicating an action. The root can be accompanied by an object a noun referring to the acting entity. The root and the object can refer to the subject, on which an action is performed. Root, subject and object therefore provide different types of information. Each can be extended with additional word dependencies. Consider the sentence Researchers conducted an experiment. The writer can focus on the object, providing more information about the researchers ("Experienced researchers conducted an experiment."), the root ("Researchers carefully conducted an experiment.") or the subject ("Researchers conducted a controlled experiment."). In a natural language, the emphasis can be expected to shift throughout the text. We hypothesize, that a deceptive writer might follow a specific strategy in providing information that could be reflected in the way he develops dependencies on the subject, root or object. Such strategy could express itself either in the average proportions of words in a sentence that are dependent on the triad, or the tendency to increase or decrease the proportion over the section of the text.

1.4.3 Author unaware of fraud

Thinking about strategic writing presupposes that the person reporting research based on data fabrication or manipulation is also aware of the fraud. However, in some cases, fraudulent papers were co-authored by a team of researchers, where only one author committed the fraud and mislead others into trusting his fabricated data. If such paper

is written up by a team member oblivious of the fraud, we cannot really assume strategic thinking. But the way of communication within the team might leave some traces on its own.

Named entities. The author responsible for fraud might be too vague of overly specific about the particulars of data collection and share too many or too few details about his work with the others. This could in turn be reflected in how specific is information provided in the article: for instance the number of references to particular organizations, or reported statistics.

Discourse markers. Even with authors unaware of fraud, there could be different levels of emotional involvement in the data. For instance, co-authors receiving fraudulent results might be more likely to display positive surprise than authors receiving non- manipulated results that are more likely to be disappointing. Therefore, the overall sentiment of the language might vary between fraudulent and non-fraudulent papers.

1.5 Design

The aim of our study is to extend the work of Markowitz and Hancock (2016) by collecting new data from 2013 onward and including more features in the classification algorithm to capture more dimensions of the complex phenomenon of language. We will operationalize the concepts discussed above and use them first on the newly collected data, then on the original dataset of Markowitz and Hancock (2016) and finally on both datasets combined. We will subsequently conduct an exploratory analysis to assess the importance of all feature sets used in our algorithms.

We hypothesize that our algorithm will achieve accuracy higher than no-information-rate (accuracy of algorithm that labels all cases as the more prevalent class), average human accuracy of 54% in deception detection (Bond Jr and DePaulo, 2006) and, as it encompasses more dimensions of language, the 57% accuracy reported by authors of the original study.

2 Methods

2.1 Raw data

We have defined fraud as a direct manipulation or fabrication of data and/or results. In some fields, important data can come from a visual source (e.g. pictures of tissue in biomedical sciences). Therefore, we have included both fabrication and manipulation of data and/or images. When an article is retracted, the journal publishes a retraction notice announcing and explaining the retraction. We have searched for papers retracted between the dates 31.12.2013 and 31.3.2018, continuing where the original study ended. The retraction notices were used to assess whether articles fulfill the following selection criteria: a) Fraud is explicitly mentioned, b) results of any related investigations are conclusive for this article (for example, when an author is convicted of fabricating data for one article, other articles of the same author can be retracted, because they are deemed unreliable. These articles would not be selected for our sample), c) the full text version of the article is available online. To represent the non-fraudulent class, each fraudulent article was matched with a non-retracted article. Given the low total prevalence of fraud, we assume that non-retracted articles are very unlikely to be undiscovered fraud. We have matched the articles on article type (e.g., research article, letter), journal, year of publication, topic derived from the abstract and keywords. The order of the matching criteria also reflects their priority. If no articles in the same year and journal matched the retraction on keywords, the same year and journal was searched for another article similar in topic. This resulted in 246 matched pairs of fraudulent and non-retracted articles. The data, code used for their extraction and final analysis are available in an OSF repository. The raw data are available upon request from the authors.

Journal	Articles
Journal of Biological Chemistry	15
PLOS ONE	10
Nature	7
Journal of Cell Science	6
Cancer Research	4
Clinical Drug Investigation	4
Journal of Management	4
PNAS	4
Science	4

Table 1: Summary of journals represented in our data. The table presents journals with at least four articles in our dataset.

2.2 Pre-processing

Because of the complex layout of scientific articles including figures, notes, headings and water-

marks, we have extracted text from the pdf files manually, creating separate .txt files for the introduction, methods, results and discussion sections of all articles. The conclusion was considered to be a part of the discussion section. When the article only had results and discussion section with no discernible separation, it was considered a results section. Subsequently, the articles were transformed from British to American English (”colour” to color) and cleared of special uses of punctuation marks (”Dr.” to ”Dr”). With the exception of readability and syntax, the features were computed from lemmatized tokens (”does” to ”do”, ”researchers” to ”researcher”).

Some authors had more than one article in our dataset. To tests whether the accuracy results from merely recognizing the same author between training and holdout set, we have assigned each article a value indicating, whether it belongs to a certain group of articles related by authors (see Figure 1). For one part of the analysis, the authors articles were split into training and holdout set, so there were no shared authorships between these two sets. For each article, a list of its authors was extracted, containing their last names and initials. The names were used to construct an adjacency matrix of a network graph with each vertex indicating an article and each edge between vertices at least one shared author. The graph was used to detect communities and attach each article a community membership value, so that every group of connected articles had the same score. This resulted in the total of 146 communities among the new articles with the average size 1.69, minimum 1 and maximum 18 (Table 2).

2.3 Features

Data were extracted using R packages Quanteda (Benoit, 2018) and SpacyR (Benoit and Matsuo, 2018). For the new dataset, we have computed all features per article section. For the old and combined dataset, features were computed per articles as a whole.

N-grams. We have extracted uni-, bi- and trigrams. The data were represented using the *bag of words* approach. Each article (section) was assigned a vector representing all n-grams in the dataset with counts for each n-gram within that article. We have selected only n-grams with sparsity across documents bellow .90. The counts were weighted using term frequency - inverse document frequency (tf-idf) weighing shown in equations 1,2 and 3. First, the frequency of each term *i* in document *j* is divided the total number of terms in document *j* (1) to compute the term frequency score (tf).

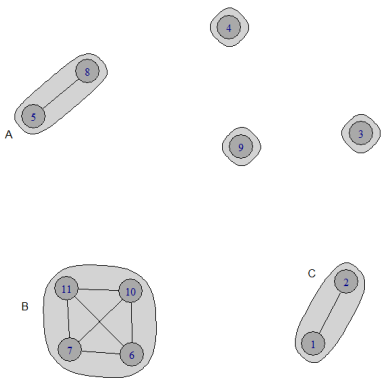

Figure 1: Small subset of our data shows shared authorship communities. Each node indicates an article, each edge at leas one shared author. A and B form communities of size 2, C constitutes a community of size 4.

Community size	Community frequency
1	111
2	18
3	4
4	3
5	3
6	2
7	1
8	3
18	1

Table 2: Community size indicates number of articles connected by shared authorships. Community size of 2 indicates two articles sharing at least one author. The right column shows the frequency of communities of various sizes in our data.

$$tf_{ij} = \frac{termcount_{ij}}{termcount_j} \quad (1)$$

Then the total number of documents and total number of documents containing term *i* are compared (2) to compute the inverse document frequency score (idf).

$$idf_i = \frac{log_{10}(documents)}{documents_i} \quad (2)$$

Tf-idf is the product of tf and idf scores (3).

$$tf\text{-}idf_{ij} = tf_{ij} idf_i \quad (3)$$

Readablity. The Flesch reading ease index was computed for each document, indicating the ease with which a text can be read, based on a) sentence length; b) word length; (3) personal word count; (4) personal sentence count (Flesch, 1948).

Hedging. Hedging was represented by quantifying certainty markers likely related to research terms. Modal or epistemic verbs and adverbs can be used to hedge against potential objections from the reader by indicating uncertainty or subjectivity (compare the words possibly and certainly, or seem and are). The varying rates of linking specific verbs and adverbs to research-related terms can be used to represent the certainty with which researchers present their work. To create a proxy for research-related terms being referred to by modal or epistemic terms, we have created a co-occurrence indexes of these terms.

First, we have selected words of interest. The modal end epistemic terms were selected from the Loughran-McDonald list from R package Lexicon (Rinker, 2018) and research by Gablasova et al. (2017), resulting in a list of words such as "maybe","actually","probably","perhaps". Research-related terms were taken from the Glossary of Research terms commissioned by the Department of Health of England (Social Care Institute for Excellence) and terms frequently observed in our data (e.g. "hypothesis", "analysis", "measurement", "model").

Then, we have computed the weighted co-occurrence of these terms in 10-word windows, so that in text "These *observations* **could** be the outcome of **some** confounding *variables* **possibly** unaccounted for in our design," the word *observation* would be considered co-occurring with words *could*, *some* and *possibly*. The co-occurrence was weighted by distance, so the co-occurrence score for *variables* and *possibly* would be higher than *observation* and *possibly*.

These scores were used to compute (weighted) relative conditional probability of any one word given any other word. First, the weighted frequency of co-occurences of words *A* and *B* in document *j* was divided by the frequency of word *B* in that document (4).

$$P(A|B)_j = \frac{Co\text{-}occurence(A,B)_j}{termcount_{Bj}} \quad (4)$$

This conditional probability was then compared with the total probability of word A occurring in the document (5).

$$WP(A|B)_j = \frac{P(A|B)_j}{P(A)_j} \quad (5)$$

Named entities. The count of recognized named entities was used to indicate the degree to which the authors were specific with information in their work. SpacyR package (Benoit and Matsuo, 2018) was used to extracted counts of named organizations, *quantity* (measurments of weight or distance), *ordinals* ("first","second") and *cardinals* (other numerals). The sum of named entities was then divided by the total numbers of words in each document.

Discourse markers. Discourse scores were computed using the LIWC lexicon (Pennebaker et al., 2015). We have used the *tone index* based on counts of words from positive and negative emotion lexicons and *analytical language index*, based on the use of categories such as prepositions, pronouns or adverbs.

Word dependencies. The informational structure of sentences was operationalized via the number of dependencies on the subject, root and object in each sentence. The dependencies were extracted using the *parse* function in SpacyR (Benoit and Matsuo, 2018). We have taken the count of direct dependencies for each of the subject-root-object triad per sentence. These were then represented as proportions of the total sentence length.

Two values per document were taken for each of the three categories. First, a mean proportion of the direct dependencies over the number of words in that sentence. Second, the slope of a linear regression model with the proportion of direct dependencies per sentence as the dependent variable and the sentence number (indicating position) as the independent variable. The slope was used to indicate the trend over time.

2.4 Analysis

R package caret (Kuhn., 2018) was used to run the analysis. First, 60% of the data was selected for a training set, ensuring proportional representation of the fraudulent and non-retracted classes. The rest of the data was kept as a holdout set. We applied 10 times repeated 10-fold cross-validation to select the optimal tuning parameters for every training.

The main analysis was done using kernel-trick support vector machine (KTSVM). A support

vector machine plots data into an n-dimensional space, with each dimension representing one feature. A hyperplane is then drawn, separating this space in a way that best separates the classes of interest (in this case, fraudulent and genuine articles). KTSVM allows for the separation of lineary non-separable data (that is, data, that could not be separated by one hyperplane in their n-dimensional space, because they do not form two separate clusters). It does so by adding additional dimensions to the existing space by computing new variables based on the existing ones. Secondary analyses were run using Random Forests (RF) and Naive Bayes (NB) algorithms.

The analysis took part in five steps. First, we have used only our newly collected data with scores on features computed per article section. Secong, we have analysed the same data, accounting for repeated authorships. Third, we have used only the articles from Markowitz and Hancock (2016). Fourth, we have used the combined dataset.

Finally, we have run an exploratory analysis to identify what sets of features most strongly contribute to the final predictive power, adding feature sets separately to the algorithm and observing the marginal increase in accuracy.

The predictive accuracy was compared to the no information rate (NIR), that is, the accuracy of a classifier that constantly labels articles as belonging to the more prevalent annotation class.

3 Results

Three machines were trained on different datasets. The NB systematically classified all articles as belonging to one class. This was most likely due to the sparsity of *n-gram*, *no stopwords n-gram* and *hedging* feature sets. In the following results, NB (but not the other machines) was always trained on data excluding the three sparse feature sets.

For each analysis, we report the detection accuracy, its 95% confidence interval, precision (the probability of correctly classifying a fraudulent article), recall (the probability that an article classified as fraudlent has been classified correctly) and area under curve (AUC), indicating the classification performance under varying decision thresholds for assigning articles to classes. We also report p-values for the comparison of accuracy to the null information rate. As we conduct multiple tests, p-values should be assesed using the bonferroni correction for all tests (k=28), so the alpha threshold should be set to .05/28≈.002.

3.1 Main analysis

Table 3 shows results for the main analysis. KT-SVM showed above chance, but not significant accuracy (Accuracy=.61, 95%CI[.54,.69], AUC=.66, p>.05). RF showed the highest and significant accuracy (Accuracy=.80, 95%CI[.73,.86], AUC=.86, p<.001). NB also performed significantly above chance (Accuracy=.75, 95%CI[.68,.81], AUC=.77, p<.001).

3.2 Accounting for shared authorship

Table 4 shows results for the main analysis when articles with the same authors did not occurr in both training and holdout dataset. KT-SVM showed above chance but (applying bonferroni correction) not significant accuracy (Accuracy=.61, 95%CI[.53,.68], AUC=.65, p>.01). RF again showed the highest and significant accuracy (Accuracy=.75, 95%CI[.68,.82], AUC=.84, p<.001). NB also performed significantly above chance (Accuracy=.68, 95%CI[.60,.75], AUC=.70, p<.001).

3.3 Old data

Table 5 shows results for articles from before 2013. KT-SVM showed above chance and significant accuracy (Accuracy=.60, 95%CI[.53,.67], AUC=.67, p=.0019). RF also showed statistically significant accuracy (Accuracy=.60, 95%CI[.5,.67], AUC=.66, p=.0019). NB did not perform significantly above chance (Accuracy=.53, 95%CI[.49,.63], AUC=.58, p¿.05).

3.4 Combined datasets

Table 6 shows results for the combined dataset. KT-SVM showed above chance but (applying bonferroni correction) not significant accuracy (Accuracy=.57, 95%CI[.52,.62], AUC=.61, p=.01). RF still showed the highest and significant accuracy (Accuracy=.61, 95%CI[.56,.65], AUC=.65, p<.001). NB did not show statistically significant performance (Accuracy=.53, 95%CI[.48,.58], AUC=.56, p=.20).

3.5 Identifying important feature sets

To identify the contribution each feature set made to the final accuracy, we have used the new data to train algorithms with step-wise addition of feature sets. We have opted for this approach as it

Machine	Accuracy	95%CI	p-Value [Acc >NIR]	Precision	Recall	AUC
KT-SVM	0.61	[0.54,0.69]	0.06	0.66	0.61	0.66
RF	0.80	[0.73,0.86]	>0.001	0.84	0.78	0.86
NB[1]	0.75	[0.68,0.81]	>0.001	0.78	0.75	0.77

Table 3: Main analysis.

Machine	Accuracy	95%CI	p-Value [Acc >NIR]	Precision	Recall	AUC
KT-SVM	0.61	[0.53,0.68]	0.01	0.61	0.62	0.65
RF	0.75	[0.68,0.82]	>0.001	0.70	0.89	0.84
NB[2]	0.68	[0.60,0.75]	>0.001	0.67	0.73	0.70

Table 4: Accounting for shared authorship.

Machine	Accuracy	95%CI	p-Value [Acc >NIR]	Precision	Recall	AUC
KT-SVM	0.60	[0.53,0.67]	0.0019	0.61	0.56	0.67
RF	0.60	[0.53,0.67]	0.0019	0.60	0.64	0.66
NB[3]	0.56	[0.49,0.63]	0.04	0.64	0.30	0.58

Table 5: Old data.

Machine	Accuracy	95%CI	p-Value [Acc >NIR]	Precision	Recall	AUC
KT-SVM	0.57	[0.52,0.62]	0.01	0.56	0.53	0.61
RF	0.61	[0.56,0.65]	>0.001	0.60	0.58	0.65
NB[4]	0.53	[0.48,0.58]	0.20	0.53	0.35	0.56

Table 6: Combined data.

shows the marginal addition of each feature set to the accuracy, while allowing for some possible interactions between the feature sets. As NB did not work with the sparse feature sets, we have run the analysis only for KT-SVM (Table 7.) and RF (Table 8.).

Early in the analysis, we have noticed that large portion of the accuracies can be attributed to *named entities*. We have therefore decided to include them as the last added feature set, to represent more clearly the role of the remaining features.

Without *named entities*, neither algorithm showed significant results. For KT-SVM, there was no difference in performance with and without *named entities* (Accuracy=.61, 95%CI[.53,.68], AUC=.65, p>.01), however, using only *named entities*, the performance increased dramatically (Accuracy=.81, 95%CI[.75,.87], AUC=.90, p<.001). RF did not show significant accuracy without *named entities* (Accuracy=.60, 95%CI[.53,.68], AUC=.67, p>.05), but performed well with *named entities* only (Accuracy=.77, 95%CI[.70,.83], AUC=.89, p<.001).

3.6 Analyzing named entities

To understand the role of *named entities*, we have compared fraudulent and non-retracted articles on their *named entity* scores per section, using the new dataset. We have run a mixed model with article sections nested within a fraudulent-non-retracted pair, testing for a main effect of an article being fraudulent and its interaction with article section. Fixed effects estimated in the model can be seen plotted in Figure 2.

We have found a significant main effect of article section (χ^2(3)=41.64, p¡.001), significant main effect of article type (χ^2(1)=87.17, p¡.001), but no significant interaction between the two (χ^2(3)=7.23, p¡.001).

4 Discussion

The aim of this study was to detect scientific fraud using linguistic cues. Our algorithms showed high performance across datasets. Despite that, the first

Features	Accuracy	95%CI	p-Value [Acc >NIR]	Precision	Recall	AUC
Syntax	0.49	[0.42,0.57]	0.94	0.57	0.32	0.51
+ LIWC	0.55	[0.47,0.63]	0.53	0.62	0.46	0.56
+ Readability	0.57	[0.49,0.64]	0.35	0.63	0.52	0.58
+ Hedging	0.56	[0.48,0.63]	0.47	0.63	0.46	0.58
+ N-grams	0.61	[0.54,0.69]	0.06	0.68	0.56	0.68
+ ns N-grams	0.61	[0.54,0.69]	0.06	0.66	0.61	0.66
+ Named entities	0.61	[0.54,0.69]	0.06	0.66	0.61	0.66
Named entities only	0.81	[0.75,0.87]	>0.001	0.83	0.84	0.90

Table 7: Stepwise SVM.

Features	Accuracy	95%CI	p-Value [Acc >NIR]	Precision	Recall	AUC
Syntax	0.52	[0.45,0.60]	0.80	0.59	0.45	0.55
+ LIWC	0.52	[0.45,0.60]	0.80	0.59	0.44	0.55
+ Readability	0.53	[0.46,0.61]	0.70	0.60	0.45	0.57
+ Hedging	0.57	[0.49,0.64]	0.35	0.66	0.44	0.58
+ N-grams	0.59	[0.51,0.66]	0.16	0.69	0.46	0.66
+ ns N-grams	0.60	[0.53,0.68]	0.10	0.68	0.53	0.67
+ Named entities	0.80	[0.73,0.86]	>0.001	0.84	0.78	0.86
Named entities only	0.77	[0.70,0.83]	>0.001	0.81	0.77	0.89

Table 8: Stepwise RF.

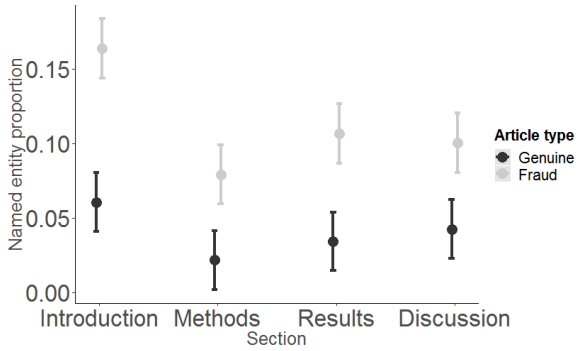

Figure 2: Estimates for named entity proportion scores are plotted per section and article type. Whiskers indicate standard errors of the estimates.

conclusion we can draw from our data is that linguistic features of academic texts contain traces of fraud to only a limited degree. Most of the observed accuracy is attributable to counts of named entities, mainly numeric values. It remains unclear, whether these values come from the number of citations, or reported statistics.

RF systematically performed better than KT-SVM. One explanation is, that KT-SVM classification was confused by the addition of irrelevant features. When using only the named entity count, KT-SVM even outperformed RF. Alternatively, the high performance of RF compared to the KT-SVM could have come from the paired nature of our data. Each fraudulent article was matched with a similar non-retracted article. When plotted in the feature space, these articles would be very likely close to each other. Then, even if the fraudulent articles were systematically in one direction relative to their non-retracted matches, it would be difficult to draw a hyperplane separating these two classes. Random Forests might be more able to distinguish between two highly similar cases.

The predictions do not generalize well in the combined dataset. There can be several reasons for that. First, the features are not section-specific, so the accuracy is generally lower. Second, the old articles cover a long period of time, ranging between years 1973-2013. Some changes in the standards of scientific publishing might be the cause of some differences between these datasets. Finally, it is possible that over time, fraud (or the capacity to detect fraud) has shifted across domains and the older data cover slightly distribution of fields or subjects than the newer data. A potential remedy for this problem would be adding the year of publication as an additional feature. However, this might pose a problem for future applications.

4.1 Practical implications

The highest obtained accuracy was around 80 % with the RF algorithm. This accuracy remained high even after accounting for multiple authorships and excluding large sets of features. While surprisingly high, it is important to consider the effect of low prior probability of fraud in any direct application. Consider an example of applying the algorithm to cancer research. Using the keyword cancer for year 2017, Web of Knowledge database returns 118328 research articles. If we assume that the proportion of researchers admitting to committing fraud reflects its true prevalence, we would state that 2% of publications are fraudulent (Fanelli, 2009). In such case, 2367 of these articles would be fraudulent and 115961. With the sensitivity and specificity being equal at .80, 1893 fraudulent articles would be marked as fraudulent (true positives) and 23192 genuine articles would be flagged as fraudulent (false positives), resulting in the precision of .075. Therefore, any such algorithm should be seen only as a screening mechanism, rather than an evidential tool.

The fact that the high accuracy resulted mainly from reported numerals makes any application further problematic. Most empirical sciences are based on quantitative analyses. Transparency is an important part of reporting an analysis and reporting statistics should not be discouraged by fraud detection algorithms.

4.2 Theoretical implications

The theoretical implications of our findings are limited, as we were not testing narrow hypotheses. Our results do not give a strong support to the notion that fraudulent scientific papers use different language than genuine publications. The named entity features consisted mostly of reported numbers. After their exclusion, semantic and syntactic features predicted fraud with only 60% accuracy and this result was not significant.

4.3 Limitations

There are several caveats to this work. First, the selection of articles was based on availability. That means the scientific fraud was represented by retracted articles. We have specifically selected articles that were undoubtedly fraudulent. However, we naturally did not have access to fraudulent articles that were not discovered. This means we might have trained our algorithm only on articles that got caught. While collecting the data, we have noticed that a part of the retractions was motivated by information from a knowledgeable informant. But even then, it is possible that we have used articles containing features that gave them away in the first place. This could also have caused the difference in named entity counts, as fraudulent articles that report more statistics might be easier to discover than those that report few. Second, we have used manual extraction to get the text files. Most scientific articles are published in the PDF format with complicated layout, page headers, subheadings, figures and notes. More importantly, retractions contain watermarks that are often very close to the original text. This made the automatic extraction impossible. While trying to ensure that no unique features leak into the copied text, we cannot entirely dismiss the possibility that some minor differences in the data occurred during the extraction process. On exploration, there were no occurrences of retraction and fraud related words in the data. Third,, the operationalization of some concepts, namely the topic-comment might have been somewhat rough and could be improved.

4.4 Future research

Future research could try to improve the predictive accuracy by fine-tuning the feature extraction, adding new dimensions and including some meta-data on publications in the analysis. More specifically, more extensive lexicons and nuanced approach could be used for the modal network, for instance by adding data on the shortest paths between its nodes. New dimensions could include non-linguistic features, for instance automatically extracted and analysed test statistics. Latent Dirichlet allocation could be used to reduce the bags of words into loadings on word-defined groups. Finally, the dataset contains articles from various disciplines and journals. Representing these in the data could add some valuable information. If the submission history of articles was available, it could be included as well.

Aside from improving algorithms, there is a potential for exploring the specific features in a more theory-driven research. We now have a unique dataset containing 499 undoubtedly fraudulent publications matched with relevant genuine articles. Besides studying potential predictors of fraud, researchers could also study the time be-

tween publication and retraction, or citation patterns of fraudulent publications.

Finally, the approach outlined in this article could be applied beyond academic fraud. Detection of HARKing (Hypothesizing After Results are Known), prediction of submission-publication gap or the citation frequency of articles are all areas, where linguistic analysis of scientific work could bring valuable insights.

5 Conclusion

We have collected scientific articles retracted for fraud and using their linguistic features, trained algorithms to detect fraudulent science. We have achieved unexpectedly high accuracy rates. Linguistic features contributed only to a small proportion of the accuracy, as most was attributable to counts of reported numerals. This article presents several areas for future research that could be of some relevance to the academic community and the public as well. We also present a unique dataset, that (if possible) will be available for further scrutiny.

References

Marcia Angell. 1986. Publish or perish: a proposal. *Annals of Internal Medicine*, 104(2):261–262.

Kenneth Benoit. 2018. *quanteda: Quantitative Analysis of Textual Data*. R package version 1.3.4.

Kenneth Benoit and Akitaka Matsuo. 2018. *spacyr: Wrapper to the 'spaCy' 'NLP' Library*. R package version 0.9.91.

Charles F Bond Jr and Bella M DePaulo. 2006. Accuracy of deception judgments. *Personality and social psychology Review*, 10(3):214–234.

Lutz Bornmann and Rüdiger Mutz. 2015. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, 66(11):2215–2222.

Chris Chambers. 2017. *The seven deadly sins of psychology: A manifesto for reforming the culture of scientific practice*. Princeton University Press.

Sacha Epskamp, Michele B Nuijten, and Sean C Rife. statcheck [online].

Daniele Fanelli. 2009. How many scientists fabricate and falsify research? a systematic review and meta-analysis of survey data. *PloS one*, 4(5):e5738.

Ferric C Fang, R Grant Steen, and Arturo Casadevall. 2012. Misconduct accounts for the majority of retracted scientific publications. *Proceedings of the National Academy of Sciences*, 109(42):17028–17033.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Santo Fortunato, Carl T. Bergstrom, Katy Börner, James A. Evans, Dirk Helbing, Staša Milojević, Alexander M. Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, Alessandro Vespignani, Ludo Waltman, Dashun Wang, and Albert-László Barabási. 2018. Science of science. *Science*, 359(6379).

Dana Gablasova, Vaclav Brezina, Tony Mcenery, and Elaine Boyd. 2017. Epistemic stance in spoken l2 english: The effect of task and speaker style. *Applied Linguistics*, 38(5):613–637.

Shuyuan Mary Ho and Jeffrey T Hancock. 2018. Computer-mediated deception: Collective language-action cues as stigmergic signals for computational intelligence. In *Proceedings of the 51st Hawaii International Conference on System Sciences*.

Shuyuan Mary Ho, Jeffrey T Hancock, Cheryl Booth, Xiuwen Liu, Muye Liu, Shashank S Timmarajus, and Mike Burmester. 2016. Real or spiel? a decision tree approach for automated detection of deceptive language-action cues. In *System Sciences (HICSS), 2016 49th Hawaii International Conference on*, pages 3706–3715. IEEE.

Bobby Lee Houtkoop, Chris Chambers, Malcolm Macleod, Dorothy VM Bishop, Thomas E Nichols, and Eric-Jan Wagenmakers. 2018. Data sharing in psychology: A survey on barriers and preconditions. *Advances in Methods and Practices in Psychological Science*, 1(1):70–85.

Ken Hyland. 1998. *Hedging in scientific research articles*, volume 54. John Benjamins Publishing.

John PA Ioannidis. 2012. Why science is not necessarily self-correcting. *Perspectives on Psychological Science*, 7(6):645–654.

Joachim Jacobs. 2001. The dimensions of topic-comment. *Linguistics*, 39(4; ISSU 374):641–682.

Ting-Ting Jiao, Ye-Min Zhang, Lin Yao, Yuan Gao, Jian Sun, Dong-Fang Zou, Guo-Ping Wu, Dan Wang, Jun Ou, and Ning Hui. 2013. Importance of spondin 1 and cellular retinoic acid binding protein 1 in the clinical diagnosis of ovarian cancer. *International journal of clinical and experimental pathology*, 6(12):3036.

Leslie K John, George Loewenstein, and Drazen Prelec. 2012. Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological science*, 23(5):524–532.

Bennett Kleinberg, Yaloe Van Der Toolen, Aldert Vrij, Arnoud Arntz, and Bruno Verschuere. 2018. Automated verbal credibility assessment of intentions: The model statement technique and predictive modeling. *Applied cognitive psychology*, 32(3):354–366.

Max Kuhn. 2018. *caret: Classification and Regression Training*. R package version 6.0-80.

Knud Lambrecht. 1996. *Information structure and sentence form: Topic, focus, and the mental representations of discourse referents*, volume 71. Cambridge university press.

Levelt Committee, Noort Committee, and Drenth Committee. november 2012. Flawed science: The fraudulent research practices of social psychologist diederik stapel.

David M Markowitz and Jeffrey T Hancock. 2014. Linguistic traces of a scientific fraud: The case of diederik stapel. *PloS one*, 9(8):e105937.

David M Markowitz and Jeffrey T Hancock. 2016. Linguistic obfuscation in fraudulent science. *Journal of Language and Social Psychology*, 35(4):435–445.

Jaume Masip, Maria Bethencourt, Guadalupe Lucas, MIRIAM SÁNCHEZ-SAN SEGUNDO, and Carmen Herrero. 2012. Deception detection from written accounts. *Scandinavian Journal of Psychology*, 53(2):103–111.

Brian A Nosek, Jeffrey R Spies, and Matt Motyl. 2012. Scientific utopia: Ii. restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, 7(6):615–631.

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

Carel F W Peeters, Chris A Klaassen, and Mark A van de Wiel. 2015. Evaluating the scientific veracity of publications by dr. jens forster.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of liwc2015. Technical report.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2017. Automatic detection of fake news. *arXiv preprint arXiv:1708.07104*.

Tyler W. Rinker. 2018. *lexicon: Lexicon Data*. Buffalo, New York. Version 1.0.0.

Uri Simonsohn. 2013. Just post it: The lesson from two cases of fabricated data detected by statistics alone. *Psychological science*, 24(10):1875–1888.

Social Care Institute for Excellence. Glossary of research terms [online].

R Grant Steen. 2011. Retractions in the scientific literature: is the incidence of research fraud increasing? *Journal of medical ethics*, 37(4):249–253.

David W Thomas, Justin Burns, John Audette, Adam Carroll, Corey Dow-Hygelund, and Michael Hay. 2016. Clinical development success rates 2006–2015. *San Diego: Biomedtracker/Washington, DC: BIO/Bend: Amplion*.

Assimakis Tseronis et al. 2009. *Qualifying standpoints. Stance adverbs as a presentational device for managing the burden of proof*. LOT, Netherlands Graduate School of Linguistics, Utrecht.

Whistleblower Report. 2012. Suspicion of scientific misconduct by dr. jens forster.

Tal Yarkoni and Jacob Westfall. 2017. Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, 12(6):1100–1122.