Utrecht University

A. J. Lindenmeyer
Kruisstraat 117
3581GK Utrecht

Classification of retracted and non-retracted scientific articles

M.Sc. Applied Data Science

A. J. Lindenmeyer (0689009)
First supervisor: Dr. Javier Garcia Bernardo
Second supervisor: Dr. Ayoub Bagheri
1st July 2022

Abstract

To retain and raise trust in science, it is essential to correct misinformation promptly, and even better to prevent the publication of incorrect information, to begin with. Taking a technical approach, this study attempts to address this critical issue of misinformation and trust in science by building models with the ability to classify retracted and non-retracted published scientific articles. These classifiers could be used by institutions to detect papers containing misinformation before they are published. Further, this study highlights the advantage of differentiating between scientific articles that have been retracted due to error and scientific articles that have been retracted due to misconduct. With this distinction, a Logistic Regression classifier was able to achieve an F1 weighted test score of 0.75 and an external validation score of 0.67.

Keywords: Retraction, scientific articles, text classification, NLP

Classification of retracted and non-retracted scientific articles

Introduction

As we see how harmful the effects of misinformation can be (e.g., in the popularity of the 'anti-vax' movement), trust in science is more important than ever. To gain this trust, researchers and institutions must follow standard research practices and publish reliable findings (D'Souza et al., 2020). Further, it is essential that violations of such are corrected, so that one can rely on the findings of published literature (D'Souza et al., 2020). Therefore, the retraction of articles is an important part of retaining the trust of both the public and researchers in science (Resnik et al., 2015).

Reasons for the retraction of a scientific article can be diverse. One of the main reasons for retraction mentioned by different studies is research misconduct (e.g., Dal-Ré, 2019). The Office of Research Integrity defines research misconduct as "fabrication, falsification, or plagiarism in proposing, performing, or reviewing research, or in reporting research results" (The Office of Research Integrity, n.d., para. 1). It further explains, "[f]abrication is making up data or results and recording or reporting them", "[f]alsification is manipulating research materials, equipment, or processes, or changing or omitting data or results such that the research is not accurately represented in the research record" and "[p]lagiarism is the appropriation of another person's ideas, processes, results, or words without giving appropriate credit" (The Office of Research Integrity, n.d., para. 2-4). In literature, research misconduct is therefore often abbreviated to FFP; fabrication, falsification, and plagiarism (DuBois et al., 2013; Resnik et al., 2015). Misconduct is a critical problem in research as it threatens public support for science and impairs the integrity of research (Resnik et al., 2015). The motivation behind research misconduct reflects flaws of the research enterprise: as success in research is often associated with the number of papers published, researchers with many publications are rewarded with status and funding, setting an incentive to publish as much as possible (Kharasch, 2021). This pressure to publish has adverse effects on the scientific integrity of researchers. Paruzel-Czachura et al. (2020) found that researchers who experience publication pressure have a higher willingness to engage in research misconduct in the future. Of course, people and institutions involved in the process of writing and publishing an article can also make mistakes, and some scientific articles are published that contain errors, and these articles must be retracted as well. The Office of Research Integrity (n.d.) specifically differentiates between research misconduct and honest error.

Incorrect information, whatever the reason for it might be, can have a multitude of negative impacts: it can pose risks to consumers or patients, decelerate medical and technological development, unnecessarily waste funds and can multiply through systematic reviews and meta-analyses (DuBois et al., 2013; Stamm, 2020). Therefore, it is critical to detect and correct such misinformation promptly. Journals have the advantage of time; to review and, if necessary, correct an article before it is published; au contraire, changing an article after publication is much more difficult (Stamm, 2020). It takes months or even years for papers to issue corrections or retractions (Stern, 2017). An automated detection system, which can notify an institution if a research paper is likely to contain misinformation prior to publication, could be a possibility to solve this problem. Allowing authors and institutions to review papers that are detected as likely to be retracted in the future, one may be able to prevent the publication of at least some papers that contain errors or are fraudulent.

Literature in the field of retracted scientific articles mainly focuses on qualitative and descriptive differences between retracted and non-retracted scientific articles; exploring different reasons for retraction (e.g., the motivation for misconduct; see Kharasch, 2021), the impact of retractions (Feng et al., 2020), temporal changes in frequency (e.g., Steen et al., 2013) or correlation of specific features with retraction (for example, number of co-authors, see Steen, 2011). Furthermore, research is often limited to a specific field (e.g., Gaudino et al., 2021). There are some studies which investigate the classification of retracted and non-retracted articles in some way. Modukuri et al. (2021) developed a model for predicting

retraction which achieved an F1 score of 71%. The authors used metadata from PubMed, a database containing abstracts and citations of biomedical literature (*PubMed about Page.*, n.d.), and from a subset, text features (e.g., p-values, sample size) and vectorized text (Abstract) as input for classification. Some studies examine the prediction of retractions based only on specific features. Copiello (2020) explored the prediction of retraction based on alternative metrics (e.g., number of views and comments), but only a fourth of retractions were predicted correctly using that approach. The SCORE program (Alipourfard et al., 2021) is taking a somewhat different approach: the goal is to compare estimates of credibility from both experts and machine algorithms. Alipourfard et al. (2021) further take evidence of reproducibility, replicability, and robustness into account to validate the provided estimates. As of now, the SCORE program has not published the results of its research.

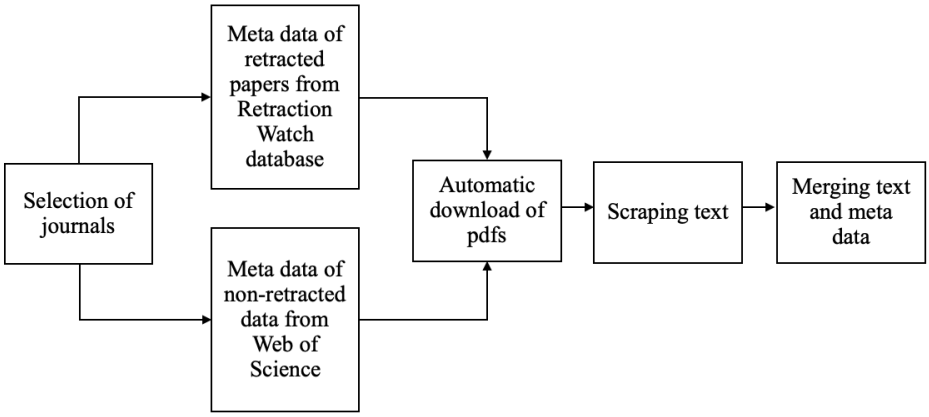
The purpose of this paper is to address this lack of quantitative approaches to the problem of scientific misinformation. The goal is to investigate if there are differences between the text of non-retracted published scientific articles and the text of retracted published scientific articles in a technical way, answering the question of if a classifier can differentiate between non-retracted scientific articles and retracted scientific articles. Moreover, as the reasons for retraction can be vastly dissimilar, we want to test if a classifier can further differentiate between scientific articles that were retracted due to error and scientific articles that were retracted due to misconduct.

Data

Non-retracted vs. retracted

Meta data of retracted scientific articles was obtained via Retraction Watch, a site devoted to scientific retractions (`http://retractiondatabase.org`). Retraction Watch is a project of the nonprofit corporation The Center for Scientific Integrity, aiding in its mission to advocate for integrity and transparency in science (The Center for Scientific Integrity, 2018). The dataset provided by Retraction Watch contains information about the *Record id, Title, Subject, Institution, Journal, Publisher, Country, Author, URLS, Article type, Retraction date, Retraction doi, Retraction PubMed id, Original paper date, Original paper doi, Original paper PubMed id, Retraction Nature, Reason, Paywalled* and *Notes* of retracted scientific articles. As it was unfeasible to access and download all retracted articles in the Retraction Watch database, 11 journals, which contained a relatively large sample of retracted papers, were selected for further use. The journals contain scientific articles about different topics, e.g., cellular biochemistry, geoscience, oncology, etc. Institutional access via Utrecht University was available for all selected journals. Because the Retraction Watch database only includes metadata of the retracted papers, we used the doi of the papers to access and download the full text. We automatically downloaded the papers included in the subset as pdfs, using the libraries 'wget' (*Wget*, n.d.), 'BeautifulSoup' (Richardson, n.d.), and 'requests' (Chandra Varanasi, 2015). To obtain scientific articles that were not retracted, we acquired metadata for non-retracted papers from the same journals that we selected before using Web of Science search results. We entered the name of the journal in *Search All*, then selected and applied the journal name as a filter of *Publication Title*. We exported the results, which we then used to download non-retracted papers automatically as pdfs. Further, we scraped the text of the downloaded pdfs using pymupdf's module 'Fitz' (*Module Fitz — PyMuPDF*, n.d.) and inserted the information into a data frame. We then merged this data frame with the meta information of both retracted and non-retracted scientific articles. An overview of the data acquisition process for both retracted and non-retracted scientific articles can be found in *Figure 1* below.

Figure 1
Data Acquisition Process



To clean the data, we removed duplicate papers. Further, we inspected if any retracted papers contained a *Retraction Notice*, which describes that a certain paper has been retracted, and if it did, removed the notice. Moreover, we removed non-retracted papers that included the words 'R/rectract/ed/ion' and 'W/withdraw/n', as this indicated that such papers might truly be retracted. Next, we removed the references of the articles by splitting the text based on the variations of the words 'References' or 'Reference List', excluding references for analysis. We discarded papers that did not contain the words 'Introduction' and 'Discussion'/'Conclusion', as those papers might not be actual scientific articles, since those typically contain the mentioned sections. Further, we removed the words 'R/rectract/ed/ion' and 'W/withdraw/n' in retracted papers, so that the classifier cannot rely on those words to differentiate between retracted and non-retracted papers. Using the spaCy library (*SpaCy · Industrial-Strength Natural Language Processing in Python*, n.d.), we applied lemmatization, stop word removal, punctuation removal, lowercasing, and removal of white spaces (and tabs). We also removed proper nouns and deleted all numbers.

After pre-processing, 6 journals contained samples from both classes. We used these 6 journals for analysis. For training and testing the classifier, we selected 4 journals (386 papers; 246 retracted, 140 non-retracted). We used the remaining 2 journals for external validation (141 papers; 72 retracted, 69 non-retracted). We picked those 2 journals as they approximately reflected the distribution of the groups in the whole dataset and because those journals contained a broad spectrum of different topics. The final datasets (2 group train/test dataset and external validation dataset), containing both retracted and non-retracted scientific articles, are shown in *Table 1* and *Table 2*.

Non-retracted vs. error vs. misconduct

For further differentiation between non-retracted scientific articles, scientific articles that were retracted due to error and scientific articles that were retracted due to misconduct, we had to use different data, as sub-selecting the retracted classes from the previous sample would have resulted in a sample too small for analysis. We obtained the data for this analysis in the same fashion. Based on the previously mentioned definition of misconduct by The Office of Research Integrity (n.d.) and the specific exclusion of errors from its definition, we included papers which stated reasons for retraction that fit either the definition of misconduct or contained the term 'error' in the misconduct or error class, respectively (see *Table 5*). For most papers, more than one reason was stated as the reason for retraction in the Retraction Watch database. Subsequently, we created a subset of the Retraction Watch database which contained the meta information of all papers which included words specified for one of the classes as the reason for retraction. 13.25% of papers from the Retraction Watch database contained a reason for retraction related to error and 30.11% of papers contained a reason for retraction related to misconduct. The most frequent reasons

for retraction were quite general or described that no specific information was available as the reason for the retraction (e.g., notice - limited or no information, investigation by journal/publisher, withdrawal).

Applying the approach of including journals in the subset that contain papers from all classes, we created a subset of the Retraction Watch database that included only papers from journals that contained more than one paper in both retracted classes (error and misconduct). Again, we downloaded the selected papers automatically as pdfs. Further, to acquire meta information of non-retracted papers, we exported results of Web of Science based on searches of the same journals (as from the retracted subset), and these non-retracted papers were downloaded automatically as pdfs as well. Again, we scraped the text of all pdfs, inserted them into a data frame and concatenated them with meta information. We applied the same pre-processing steps as mentioned above to the data. Additionally, as there were mainly multiple reasons as the cause for retraction stated in the Retraction Watch database, we removed papers which included reasons from both the misconduct and error class. After pre-processing, the sample contained 1484 papers (158 error, 385 misconduct, 941 non-retracted) from 42 journals. We split this sample into two subsets; one for training and testing the classifier, consisting of 40 journals (1117 papers; 119 error, 306 misconduct, 692 non-retracted) and the other for external validation, consisting of 2 journals (67 papers; 39 error, 79 misconduct, 249 non-retracted). We picked those journals for external validation as they reflected the distribution of the classes in the whole dataset. The final datasets (3 group train/test dataset and external validation dataset), containing non-retracted scientific articles and scientific articles retracted due to error and misconduct, are shown in *Table 3* and *Table 4*.

Ethical considerations

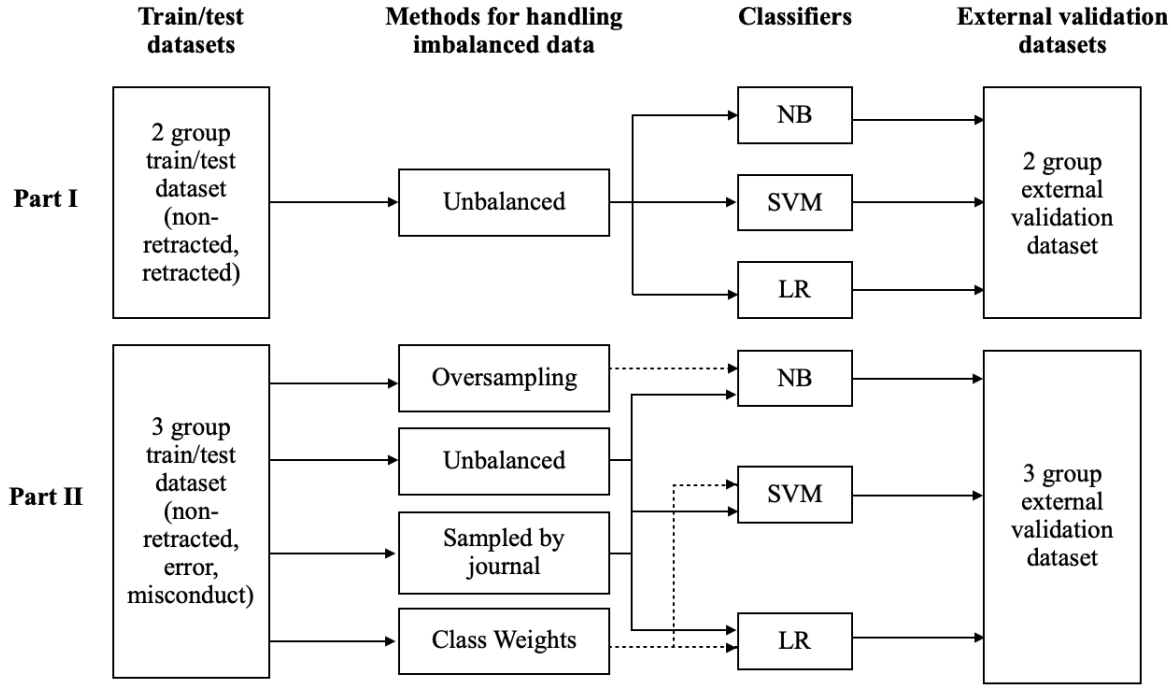
Data [was made] available from The Center For Scientific Integrity, the parent nonprofit organization of Retraction Watch, subject to a standard data use agreement (The Center for Scientific Integrity, 2018. para. 5). Due to legal reasons, the Retracted Watch database cannot be published. Ethical standards were met since no author names or names of affiliated persons of retracted papers were used for analysis or were published, and thus no harm was inflicted.

Methods

To answer the question if a classifier can differentiate between non-retracted and retracted scientific articles, and further between non-retracted scientific articles, articles retracted due to error and articles retracted due to misconduct, we performed two analyses.

In the first part of this analysis, we tested how different models perform in the classification task of differentiating between non-retracted and retracted scientific articles, using the 2 group train/test dataset and the 2 group external validation dataset. In the second part, we tested if models can further differentiate the retracted scientific articles, based on the reason of retraction (error/misconduct), using the 3 group train/test dataset and the 3 group external validation dataset. *Figure 2* depicts the methods for handling imbalanced data and classifiers used for analysis parts I and II.

Figure 2
Methods and classifiers used for analysis



Dottet lines represent sampling methods not used for all classifiers.

Classifiers

We chose Naive Bayes (NB), Support Vector Machine (SVM) and Logistic regression (LR) as models for the classification task. We selected NB as it is a simple model that utilizes the Bayes rule and assumes conditional independence of attributes (Webb, 2010). It can be used as a benchmark to compare other models to, and despite its simplicity, often performs relatively well (Webb, 2010). Multinomial NB is a variant that is typically used for text classification tasks and is mentioned to work well with tf-idf vectors (Pedregosa et al., 2011). SVM is another model that has proven to perform well in the context of text classification (Sun et al., 2009; Z. Liu et al., 2010). Based on the margin maximization principle, SVM solves a classification task by creating the hyperplane which separates the classes in the most favourable way (Adankon & Cheriet, 2009). Some papers further mention Logistic Regression as a well-performing model in text classification tasks (e.g., Pranckevičius & Marcinkevičius, 2017). In our case, LR estimates the probability of an article belonging to one of the classes and assigns the article to the most likely class.

Performance measures

As a measure of performance, the weighted F1 score is a suitable metric as it takes the weighted average of the F1 scores of the different classes, taking the class imbalance in the data used for the second part of the analysis into account. Further, recall and precision can reveal how the classifiers perform for different classes. Recall reflects a classifier's ability to successfully identify cases that belong to a certain class. Precision, on the other hand, shows if a classifier is able to not assign a case to a class to which it does not belong to. For all measures, the lowest value achievable is 0 and the highest is 1. The formulas for the F1 score of one class, recall and precision are shown below.

$$F1_{class} = 2 * (precision_{class} * recall_{class})/(precision_{class} + recall_{class})$$

$$recall = true\,positives/(true\,positives + false\,negatives)$$

$$precision = true\,positives/(true\,positives + false\,positives)$$

We used F1 weighted to compare the overall performance of the classifiers, as it considers both precision and recall of all classes. Although the weighting of F1 scores might favour the majority class (for the 3 group dataset, the non-retracted class), this measure corresponds to the goal of building a prototype for an automated detection system which can be implemented in the real world. Even though some cases of papers containing misinformation might be missed by the classifiers in this way, emphasising the minority classes would lead to an increase in false positives for those classes (e.g., classification into likely to be retracted due to error or misconduct, but the paper contains no misinformation). This in turn would result in a high workload for institutions or journals reviewing those flagged papers, which would make the implementation of such a model in practice unfeasible.

Vectorization

The selected models require numeric features as an input; thus, to execute the analysis, the text of the scientific articles needed to be transformed into vector representations (Bengfort et al., n.d.). Term frequency–inverse document frequency is a favourable way to vectorize text, since this bag-of-words representation encodes the normalized frequency of words in an article with respect to the frequency in other articles, emphasizing words that are relevant for a particular article while taking the context of the article, the corpus (i.e., all articles), into account (Bengfort et al., n.d.).

$$tfidf\ (term,\ document,\ corpus) = tf\ (term,\ document)\ *\ \ idf\ (term,\ corpus)$$

After converting the pre-processed collection of articles into tf-idf features using scikit-learn's TfidfVectorizer (Pedregosa et al., 2011), the models were trained and tested. We first trained and tested the models using the 2 group train/test dataset, and then externally validated the models using the 2 group external validation dataset. For the second part of the analysis, we repeated this procedure using the 3 group train/test dataset and the external validation dataset.

Handling imbalanced data

For the second part of the analysis, the issue arose that the groups error, misconduct and non-retracted varied significantly in size. Different strategies were used to address this problem of class imbalance. All classifiers were run with the unbalanced sample, to use as a comparison point to evaluate if strategies to balance the classes also led to better test performance and higher scores for external validation. Further, for SVM and LR, the parameter 'class weight' could be specified as 'balanced'. The parameter differently penalizes a false classification of the minority and the majority, in a manner that the applied weight is inversely proportional to the frequency of the class (Lemaıtre, Nogueira, et al., 2017). For NB, this option is not available. Thus, imbalanced learn's RandomOversampler was used to balance the different groups (Lemaıtre, Nogueira, et al., 2017). Finally, a sample was created that contained the same number of articles for each class and journal, subsequently referred to as 'sampled by journal', which was then used as an input for all classifiers.

Hyperparameter tuning

To find optimal parameter settings for SVM and LR, hyperparameter tuning was used. F1 weighted was used as a scoring measure. For SVM, the optimal settings for the parameters C, kernel and gamma were determined in hyperparameter tuning, for LR, the optimal settings for C, solver and penalty were determined. For analysing the 2 group train/test dataset, the optimal parameter settings were detected to be C=1.0, kernel=linear and gamma='scale' for SVM, and C=100, solver='newton-cg' and penalty='l2' for LR. For analysing the 3 group train/test dataset, optimal settings for the parameters depended on the training sample. For SVM, using an unbalanced data sample with or without specified class weights, the best settings were discovered to be C=10, kernel='sigmoid', gamma='scale'. For analysing the sample which was 'sampled by journal', C=100, kernel='rbf', gamma='scale' were the favourable settings to use. For LR, using an unbalanced data sample with or without class weights, C=100, solver='liblinear', and penalty="l2" were found to be the best settings for the parameters. When using a sample with balanced classes sampled by journal, the settings solver='newton-cg', penalty="l2", C=100 produced the best F1 weighted score for LR.

Model input

For the first part of the analysis, the 2 group train/test and external validation datasets were used as input into the classifiers (NB, SVM, LR). The performance of the classifiers was measured on a test set split off from the 2 group train/test dataset, and external validation was tested on the 2 group external validation dataset. For the second part of the analysis, a sample from the 3 group train/test dataset was used to fit the classifiers. To handle the imbalanced classes, different samples of the 3 group train/test dataset were used to train the classifiers as described in the section *Handling imbalanced data* above. The classifiers' performance was then measured on a test set split off from the 3 group train/test dataset, and external validation was tested on the 3 group external validation dataset.

Results

F1 weighted was used to evaluate the overall performance of the classifiers. Further, recall and precision were used as a measure to evaluate the models' performance to classify specifically the different classes.

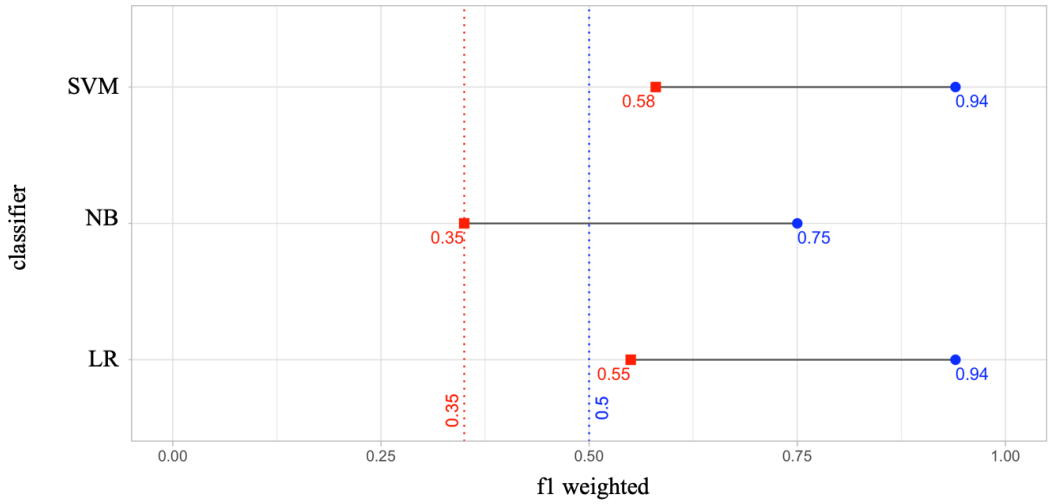
Analysis part I

F1 weighted testing and external validation scores of NB, SVM and LR are visualized in *Figure 3*. Results of a dummy classifier were used as a reference point for each of the F1 weighted test and the F1 weighted external validation scores, using the 2 group train/test dataset and external validation dataset, respectively. SVM and LR performed equally well regarding their F1 weighted test scores, both having a score of 0.94. In external validation, SVM was able to reach a higher F1 weighted score (0.58) compared to LR (0.55). From the 97 scientific articles used for testing, SVM was able to correctly classify 91; in external validation, SVM could only classify 84 out of 141 scientific articles correctly. In comparison to SVM and LR, NB performed worse concerning its F1 weighted test score (0.75) and F1 external validation score (0.35). NB reached a higher F1 weighted test score (0.75) than a dummy classifier (0.5) but performed just as well as a dummy classifier in external validation (0.35). SVM and LR were both able to reach a higher F1 weighted test and external validation score compared to the dummy classifier.

Even though SVM and LR had very high F1 weighted scores for testing, the scores dropped substantially in external validation. Looking at the recall, precision and F1 scores per class of the SVM classifier (see *Figure 4*), the model performs worse in classifying both classes in external validation compared to testing. Especially the recall score of the non-retracted class is significantly lower in external validation. This means that in external validation, the SVM classifier tends to assign the articles to the retracted class. LR exhibits this same tendency to classify articles as retracted in external validation. NB already shows

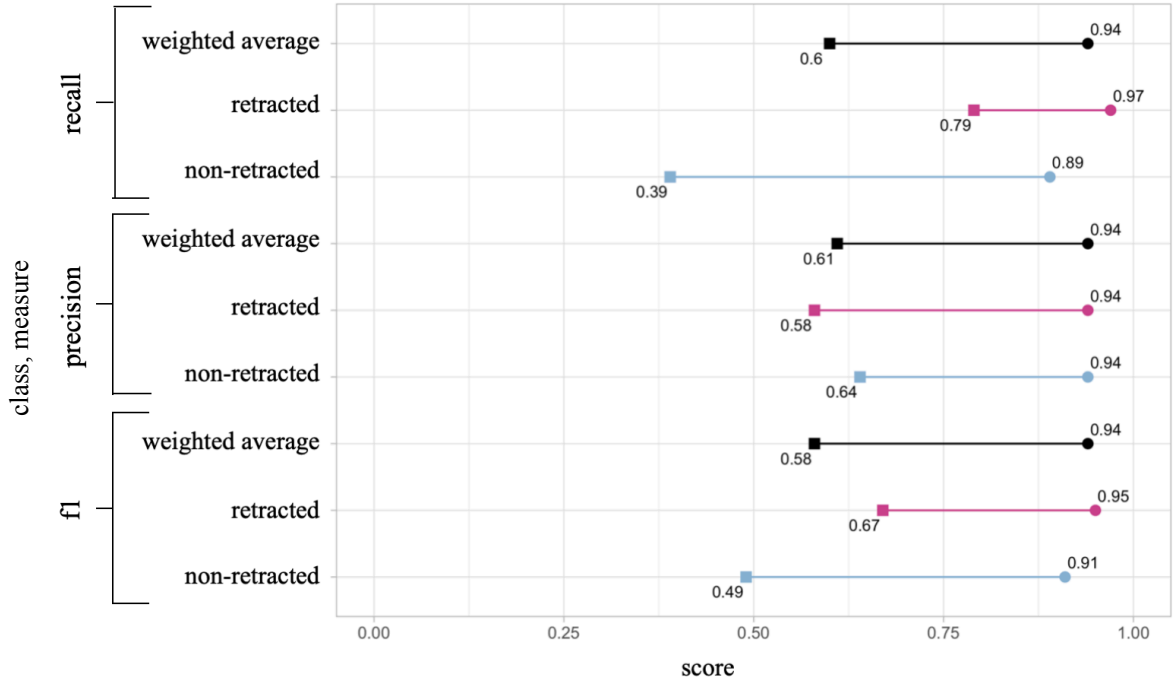
a bias towards classifying articles as retracted in testing; in external validation, all articles were classified as retracted. For SVM, both precision and F1 scores were similar for both groups in testing and external validation. A summary of F1, recall and precision scores for the different classifiers and classes can be found in *Table 6*.

Figure 3
F1 weighted scores of NB, SVM and LR using the 2 group train/test dataset and external validation dataset



Blue points represent F1 weighted test scores using the 2 group train/test dataset, red squares represent F1 weighted scores for external validation using the 2 external validation dataset. The blue dotted line is the F1 weighted test score of a random classifier. The red dotted line indicates the F1 external validation score of a dummy classifier.

Figure 4
SVM F1, recall and precision scores for the classes retracted and non-retracted and weighted average, using the 2 group train/test dataset and external validation dataset

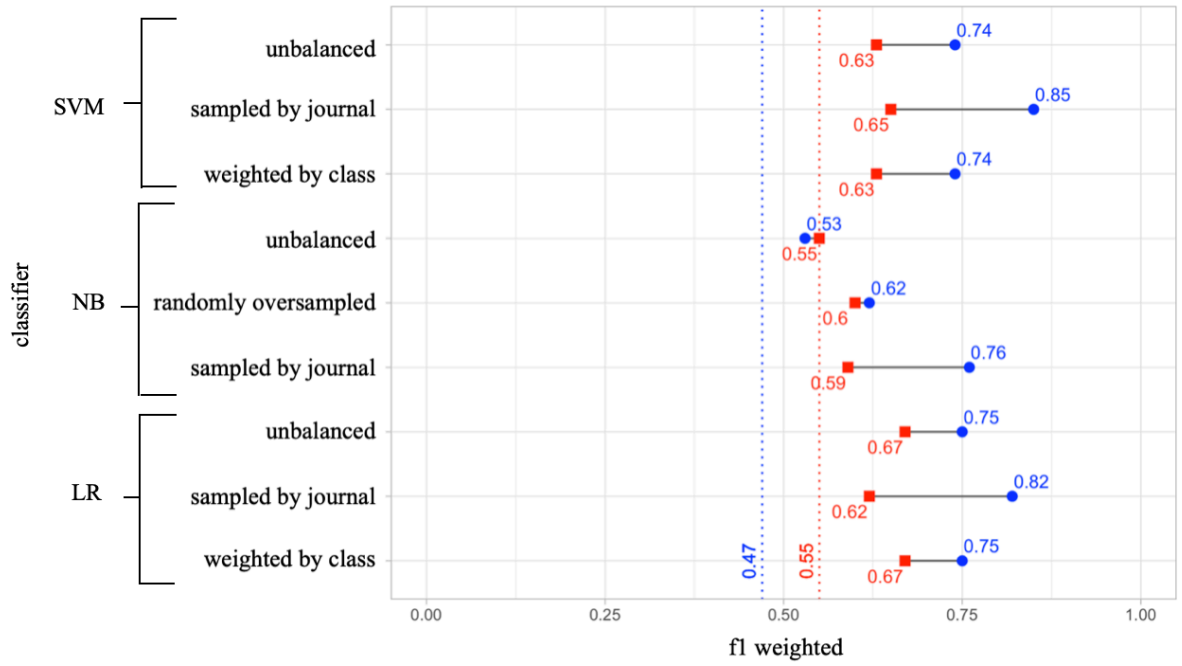


Points represent test scores using the 2 group train/test dataset, squares represent scores for external validation using the 2 group external validation dataset. Black represents the weighted average of the classes of the specific measure, pink represents the retracted and light blue represents the non-retracted class.

Analysis part II

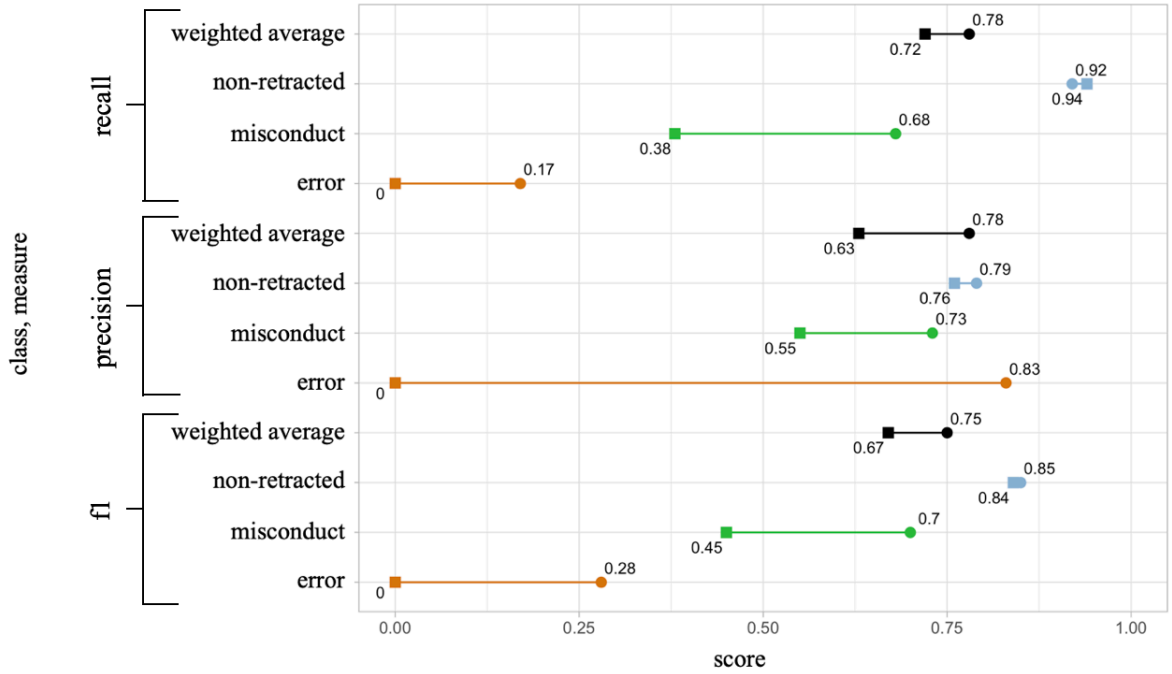
The test and external validation F1 weighted scores of the classifiers (NB, SVM, LR) are shown in *Figure 5*. Again, dummy classifier results were used as a reference point for each F1 weighted test (using the 3 group train/test dataset) and the external validation (using the 3 group external validation dataset) scores. All classifiers, except NB with an unbalanced sample as input, had higher F1 scores for testing and external validation than the dummy classifier. SVM, using data with balanced classes sampled by journal, produced the highest F1 weighted score of 0.85. However, the F1 weighted score of external validation (0.65) fell short compared to other classifiers. Generally, all classifiers reached their highest F1 weighted testing scores using this method but also perform considerably worse in external validation. The highest F1 weighted external validation score of 0.67 was achieved by LR, both with an unbalanced dataset and with the same dataset but additionally specifying class weights to be balanced. The LR classifier also produced the same F1 weighted testing score of 0.75 for both methods. With the unbalanced sample, LR was able to correctly classify 217 out of 280 articles in testing, and 265 out of 367 in external validation. With specifying class weights to be balanced, LR correctly classified 216 out of 280 articles in testing, and 260 out of 367 in external validation. Further, we investigated LR's ability to distinguish the classes using an unbalanced sample or specifying class weights (see *Figure 6* and *Figure 7*, respectively). For both methods, recall scores were highest for non-retracted, the largest class, and lowest for error, the smallest class, both in testing and external validation. Precision scores for both methods were similar for testing, but for external validation, the same pattern as in recall emerged; the classifier performed best for the largest class, non-retracted, and worst for the smallest, error. For both methods, LR shows a tendency to classify articles to the non-retracted class, to a greater extent in external validation compared to testing. A summary of F1, recall and precision scores for the different classifiers, methods and classes can be found in *Table 7*.

Figure 5
F1 weighted scores of NB, SVM and LR using no and different sampling methods and the 3 group train/test and external validation datasets



Blue points represent F1 weighted test scores using the 3 group train/test dataset, red squares represent F1 weighted scores for external validation using the 3 group external validation dataset. The blue dotted line is the F1 weighted test score of a random classifier. The red dotted line indicates the F1 external validation score of a dummy classifier.

Figure 6
LR unbalanced, F1, recall and precision scores for the classes error, misconduct, non-retracted and weighted average, using the 3 group train/test and external validation datasets



Points represent test scores using the 3 group train/test dataset, squares represent scores for external validation using the 3 group external validation dataset. Black represents the weighted average of the classes of the specific measure, light blue represents the non-retracted, green the misconduct and orange the error class.

Discussion

To answer the initial part of the research question, we tested whether classifiers (NB, SVM and LR) can distinguish the text of retracted scientific articles from the text of non-retracted scientific articles. Performances of SVM and LR (measured by their F1 weighted testing and external validation scores) using the 2 group train/test dataset and external validation dataset exceeded the performance of a dummy classifier. However, NB performed relatively well in testing but just as well as a dummy classifier in external validation. Therefore, SVM and LR were able to differentiate between the text of retracted and non-retracted articles. For NB, this does not seem to be the case. LR and SVM performed equally well in testing, but SVM was able to reach the highest F1 weighted score for external validation for the 2 group dataset. Important to note is that the difference between the F1 weighted test score and the F1 weighted external validation score was 36-40% for all classifiers, thus, they all achieved a considerably lower score in external validation.

Next, we used the 3 group train/test dataset and external validation dataset as inputs to NB, SVM and LR to investigate if a classifier can further differentiate between non-retracted scientific articles, scientific articles that were retracted due to error and scientific articles that were retracted due to misconduct. Different methods were used to handle the problem of imbalanced classes, as mentioned above. All classifiers using different methods were able to produce higher F1 testing and external validation scores than a dummy classifier, except NB trained on unbalanced data. Consequently, both SVM and LR were able to distinguish between non-retracted articles, articles retracted due to error and articles retracted due to misconduct. NB was only able to differentiate between these three classes if the input dataset used for training was adjusted to balance the classes in some way. Except for NB with unbalanced data as input, the classifiers using the 3 group datasets as inputs also exhibited the pattern of higher F1

weighted test scores compared to F1 weighted external validation scores, though the differences were not as extreme (here, 2-20% for all classifiers and methods). The classifiers reached their highest F1 weighted test scores when data imbalance was handled with 'sampling by journal', but this also led to a more severe drop in performance in external validation, compared to the other methods. External validation can be argued to be the more critical measure for the performance evaluation of classifiers. Models used for prediction often reach higher performance scores when testing on data which was also used for training (from the same data pool), possibly leading to an overestimation of the model's performance (Bleeker et al., 2003). Most importantly, external validation is essential to be able to generalize to other, new samples and reproduce found results (Ramspek et al., 2021). Examining the classifiers' performance in external validation, LR was able to reach the overall highest F1 weighted score, both with unbalanced data and with specified class weights. This further suggests that trying to handle class imbalance did not improve the performance of LR when testing on external data.

The classifiers' different performance for the classes reflects the imbalance of non-retracted articles, retracted articles due to error and retracted articles due to misconduct. The difference in the sample size of articles retracted due to error vs. due to misconduct is apparent in the Retraction Watch database, as 13.25% of articles contain a reason for retraction related to error and 30.11% to misconduct. In a sample of retracted articles used by Dal-Ré (2019), research misconduct (FFP) was the reason for 53% of retractions, and error for 21%; other frequent reasons listed for retraction were duplicate publication or limited / no information. Ideally, the external validation dataset should reflect the distribution of the classes in the real world. Although studies (e.g., Dal-Ré 2019) show a similar distribution of retracted articles due to error and retracted articles due to misconduct, it is impossible to know what the actual distribution looks like. Further, the proportion of retracted and non-retracted articles remains unknown. Presumably, numerous scientific articles containing misinformation will never be disclosed as such, and thus will never be retracted, even though they truly should.

Investigating the models' features that are used to differentiate between the classes, both for the first and second parts of the analysis, the models seem to mainly rely on topic-specific words for classification (*Table 8* shows the top 30 indicative words used by the models). This explains the inferior performance of the classifiers (except NB fit with the unbalanced 3 group train/test dataset) on new, unseen data in external validation compared to their performance on test data from the same journals they were trained on. Both external validation datasets contain scientific articles from journals on which the classifier has not been trained. Thus, the classifiers, which rely heavily on topic- or journal-specific words that might indicate if a paper is retracted (due to some error/misconduct) or not, perform worse when those words are not present in the external data. Therefore, the classifiers which perform better on the test data than on the external data likely overfit the training data. Furthermore, some journals about specific topics (e.g., Journal of Cellular Biochemistry) contain more scientific articles from some classes and fewer from others. This is true for the datasets used for either part of the analysis. Consequently, some topic words are indicative of the classes that contain more samples related to certain topics. In journal-based sampling, to balance the classes, the problem is addressed as a sample with equal class distribution for each journal is used as input for the classifiers. However, a different problem arises; since papers can be included in the training set multiple times, some topic-related words that are contained in these duplicate papers could be very indicative of the minority classes. Based on the relatively large difference in testing and external validation F1 weighted scores when using this method, this method seems to mainly lead to overfitting the training data. Though classifiers using different methods all seem to rely on topic- or journal-related words to differentiate between classes, this is not necessarily a problem. In the real world, some journals and topics also have a higher percentage of retracted articles - or contain more articles retracted due to misconduct than error, so a classifier based on topic-related words can still be useful. To improve the models' performance for predicting if certain scientific articles are likely to be retracted (due to some reason), it could be beneficial to train the classifiers on journals that are relevant for the predictions. Training the classifiers on journals that are used in testing and training on a variety of journals covering different topics are adequate approaches, favourable would be using both. Different studies suggest a relation between both retraction and journals and retraction and topic. Fang & Casadevall (2011) found that the frequency of retraction differs between journals and is strongly related to the journal impact factor. Retraction is especially common in biomedical and multidisciplinary

journals, while social sciences, arts, and humanities have significantly lower retraction rates (Lu et al., 2013). This uneven distribution of topics is also reflected in the Retraction Watch database.

Compared to the models' performance with retracted and non-retracted classes as input, testing performance was overall lower when using three classes. However, F1 scores of the models were generally higher when using three classes as input. Considering external validation to be the most important measure, it can be concluded that the further distinction of retracted scientific articles into articles that are retracted due to error and articles that are retracted due to misconduct can even improve the models' performance, as all models (except for NB trained on unbalanced data) were able to achieve higher F1 weighted external validation scores in the second part of the analysis.

An institution could put this classifier to practice by further training the classifier on more articles (both non-retracted and retracted due to error/misconduct), especially articles from journals that it should be tested on, and/or journals covering different topics. Newly submitted articles that are reviewed by the institution for publication can be used as input for the classifier, which assigns the new article to the most likely class. Applying this approach, employees of the institution can focus and spend more time on reviewing papers classified as likely to be retracted. Ideally, they can work more efficiently and identify articles containing misinformation before they are published.

Limitations

Although the class imbalance of the 3 group train/test dataset might be related to an imbalance of retracted (due to error/misconduct) and non-retracted scientific articles in the real world, we do not know the exact distribution of retracted papers (due to error/misconduct) in the real world. Furthermore, the imbalance generally led to lower recall, precision, and F1 scores of minority classes (misconduct and error). Therefore, this issue had adverse effects on the performance of the classifiers regarding accurately predicting the retraction likelihood for all scientific articles, and on a larger scale also negatively impacts the goal of being able to successfully identify retracted research articles. Additionally, the performance of the classifier depends on the training and testing data, thus the selection of journals used for the train/test and external validation datasets affect their performance. Another limitation of this study is that the selection of the classes error and misconduct and further allocation of certain reasons for retraction in the Retraction Watch database to those classes, though based on definitions, remains subjective to at least some degree. Moreover, as the number of papers available for the classes error and misconduct was limited, no time frame for the sample papers of the different groups was set. This issue was addressed with the removal of numbers in text in pre-processing, for the years not to be indicative of the classes. Additionally, the split into different sections (Abstract, Introduction, Method, Discussion/Conclusion and References), may have not always been correct. Unfortunately, checking if every paper is split correctly is not feasible. Consequently, some part of the text of an article or the whole article (as papers without these sections were removed from the datasets) might have been wrongfully discarded.

Future research

Future studies could address the imbalance of classes by trying to retrieve more retracted scientific articles due to both error and misconduct, e.g., from other databases. Furthermore, it could be tested if the classifiers can be improved by adding other features than text to the models for prediction. For example, the program SCORE plans to include results from semantic analysis (e.g., sample sizes, methods, experimental variables) to produce machine-generated estimates of credibility (Alipourfard et al., 2021). In addition, metadata (e.g., date of publication, institution, publisher, etc.) could be added to the model. However, one must be careful to not include sensitive data in models, e.g., author names should not be indicative of a paper to be retracted. Further, one must be careful not to make accusations; no model classifies every input correctly, and authors or other affiliated persons should not face negative consequences based on the result of an algorithm.

Conclusion

In this study, we took a technical approach to address the problem of retractions of scientific articles, and the broader problem of scientific misinformation. The aim was to determine if a classifier could distinguish between non-retracted and retracted scientific articles, and further between articles that were retracted due to error and due to misconduct, based on text. This goal was reached to some degree; an LR model, with non–retracted scientific articles and scientific articles retracted due to error and misconduct (except references) used as input, performed well both in testing and external validation regarding the weighted F1 scores (0.75 and 0.67, respectively). Especially for external validation, the classifier was not able to reach high scores for recall, precision and F1 for the minority classes, scientific articles retracted due to misconduct and error. Training the model on journals that it will be tested on or training it on a variety of journals covering different topics could be profitable and improve the model's performance in allocating input papers to the correct classes, especially minority classes. This study contributes to existing research, shining a light on possible solutions for catching erroneous or fraudulent articles before they are published and need to be retracted. Reducing the number of published papers containing misinformation also means promoting trust in science.

Acknowledgements

I would like to thank my fellow students Joseph Johan Ignace Franssen and Eveline Schmidt for their contribution to this thesis. Further, I am grateful for the support and insightful comments of my supervisors Dr. Javier Garcia Bernardo and Dr. Ayoub Bagheri during all stages of the thesis.

References

Adankon, M., Mathias M.and Cheriet. (2009). Support vector machine. In A. Li Stan Z.and Jain (Ed.), *Encyclopedia of biometrics* (pp. 1303–1308). Boston, MA: Springer US. Retrieved from https://doi.org/10.1007/978-0-387-73003-5$_2$99 doi: 10.1007/978-0-387-73003-5$_2$99

Alipourfard, N., Arendt, B., Benjamin, D. M., Benkler, N., Bishop, M. M., Burstein, M., ... Wu, J. (2021). Systematizing confidence in open research and evidence (SCORE). Retrieved from https://doi.org/10.31235/osf.io/46mnb doi: 10.31235/osf.io/46mnb

Benjamin, B., Rebecca, B., & Tony, O. (n.d.). *Text vectorization and transformation pipelines—applied text analysis with python. chapter 4. text vectorization and transformation pipelines.* https://www.oreilly.com/library/view/applied-text-analysis/9781491963036/ch04.html. (Accessed: 2022-06-22)

Bleeker, S., Moll, H., Steyerberg, E., Donders, A., Derksen-Lubsen, G., Grobbee, D., & Moons, K. (2003). External validation is necessary in prediction research:. *Journal of Clinical Epidemiology*, *56*(9), 826–832. Retrieved from https://doi.org/10.1016/s0895-4356(03)00207-5 doi: 10.1016/s0895-4356(03)00207-5

Chandra, R. V., & Varanasi, B. S. (n.d.). *Python requests essentials.* Packt Publishing Ltd.

Copiello, S. (2020). Other than detecting impact in advance, alternative metrics could act as early warning signs of retractions: tentative findings of a study into the papers retracted by PLoS ONE. *Scientometrics*, *125*(3), 2449–2469. Retrieved from https://doi.org/10.1007/s11192-020-03698-w doi: 10.1007/s11192-020-03698-w

Dal-Ré, R. (2019). Analysis of retracted articles on medicines administered to humans. *British Journal of Clinical Pharmacology*, *85*(9), 2179–2181. Retrieved from https://doi.org/10.1111/bcp.14021 doi: 10.1111/bcp.14021

Definition of research misconduct. (n.d.). ORI - The Office of Research Integrity. Retrieved from https://ori.hhs.gov/definition-research-misconduct (Accessed: 2022-06-22)

D'Souza, D. M., Sade, R. M., & Moffatt-Bruce, S. D. (2020). The many facets of research integrity: What can we do to ensure it? *The Journal of Thoracic and Cardiovascular Surgery*, *160*(3), 730–733. doi: 10.1016/j.jtcvs.2019.12.127

DuBois, J. M., Anderson, E. E., Chibnall, J., Carroll, K., Gibb, T., Ogbuka, C., & Rubbelke, T. (2013). Understanding research misconduct: A comparative analysis of 120 cases of professional wrongdoing. *Accountability in Research*, *20*(5-6), 320–338. Retrieved from https://doi.org/10.1080/08989621.2013.822248 doi: 10.1080/08989621.2013.822248

Fang, F. C., & Casadevall, A. (2011). Retracted science and the retraction index. *Infection and Immunity*, *79*(10), 3855–3859. Retrieved from https://doi.org/10.1128/iai.05661-11 doi: 10.1128/iai.05661-11

Feng, L., Yuan, J., & Yang, L. (2020). An observation framework for retracted publications in multiple dimensions. *Scientometrics*, *125*(2), 1445–1457. Retrieved from https://doi.org/10.1007/s11192-020-03702-3 doi: 10.1007/s11192-020-03702-3

for Scientific Integrity, T. C. (n.d.). *Retraction watch database.* Retrieved 2022-06-22, from https://retractionwatch.com/retraction-watch-database-user-guide/

Gaudino, M., Robinson, N. B., Audisio, K., Rahouma, M., Benedetto, U., Kurlansky, P., & Fremes, S. E. (2021). Trends and characteristics of retracted articles in the biomedical literature, 1971 to 2020. *JAMA Internal Medicine*, *181*(8), 1118. Retrieved from https://doi.org/10.1001/jamainternmed.2021.1807 doi: 10.1001/jamainternmed.2021.1807

Kharasch, E. D. (2021). Scientific integrity and misconduct—yet again. *Anesthesiology*, *135*(3), 377–379. Retrieved from https://doi.org/10.1097/aln.0000000000003916 doi: 10.1097/aln.0000000000003916

Lemaitre, G., Nogueira, F., & Aridas, C. K. (2016). *Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning.* arXiv. Retrieved from https://arxiv.org/abs/1609.06570 doi: 10.48550/ARXIV.1609.06570

Liu, Z., Lv, X., Liu, K., & Shi, S. (2010). Study on SVM compared with the other text classification methods. In *2010 second international workshop on education technology and computer science.* IEEE. Retrieved from https://doi.org/10.1109/etcs.2010.248 doi: 10.1109/etcs.2010.248

Lu, S. F., Jin, G. Z., Uzzi, B., & Jones, B. (2013). The retraction penalty: Evidence from the web of science. *Scientific Reports*, *3*(1). Retrieved from https://doi.org/10.1038/srep03146 doi: 10.1038/srep03146

Modukuri, S. A., Rajtmajer, S., Squicciarini, A. C., Wu, J., & Giles, C. L. (2021). *Understanding and predicting retractions of published work* (Unpublished doctoral dissertation).

Module fitz — PyMuPDF 1.20.0 documentation. (n.d.). Retrieved 2022-06-26, from https://pymupdf.readthedocs.io/en/latest/module.html

Paruzel-Czachura, M., Baran, L., & Spendel, Z. (2020, December). Publish or be ethical? publishing pressure and scientific misconduct in research. *Research Ethics*, *17*(3), 375–397. Retrieved from https://doi.org/10.1177/1747016120980562 doi: 10.1177/1747016120980562

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, (2011). Scikit-learn: Machine learning in python. Retrieved from https://arxiv.org/abs/1201.0490 doi: 10.48550/ARXIV.1201.0490

Pranckevičius, T., & Marcinkevičius, V. (2017). Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification. *Baltic Journal of Modern Computing*, *5*(2). Retrieved from https://doi.org/10.22364/bjmc.2017.5.2.05 doi: 10.22364/bjmc.2017.5.2.05

Ramspek, C. L., Jager, K. J., Dekker, F. W., Zoccali, C., & van Diepen, M. (2021). External validation of prognostic models: what, why, how, when and where? *Clinical Kidney Journal*, *14*(1), 49–58. Retrieved from https://doi.org/10.1093/ckj/sfaa188 doi: 10.1093/ckj/sfaa188

Resnik, D. B., Neal, T., Raymond, A., & Kissling, G. E. (2015). Research misconduct definitions adopted by u.s. research institutions. *Accountability in Research*, *22*(1), 14–21. Retrieved from https://doi.org/10.1080/08989621.2014.891943 doi: 10.1080/08989621.2014.891943

Retraction watch database. (2018). Retraction Watch. Retrieved from http://retractiondatabase.org

Richardson, L. (n.d.). *Beautiful soup documentation.*

spaCy · industrial-strength natural language processing in python. (n.d.). Retrieved 2022-06-27, from https://spacy.io/

Stamm, T. (2020). From honest mistakes to fake news – approaches to correcting the scientific literature. *Head &amp Face Medicine*, *16*(1). Retrieved from https://doi.org/10.1186/s13005-020-00220-8 doi: 10.1186/s13005-020-00220-8

Steen, R. G. (2011, November). Retractions in the scientific literature: do authors deliberately commit research fraud? *Journal of Medical Ethics*, *37*(2), 113–117. Retrieved from https://doi.org/10.1136/jme.2010.038125 doi: 10.1136/jme.2010.038125

Steen, R. G., Casadevall, A., & Fang, F. C. (2013). Why has the number of scientific retractions increased? *PLoS ONE*, *8*(7), e68397. Retrieved from https://doi.org/10.1371/journal.pone.0068397 doi: 10.1371/journal.pone.0068397

Stern, V. (2017). *The retraction countdown: How quickly do journals pull papers?* Retraction Watch. Retrieved from shorturl.at/bfkwF

Sun, A., Lim, E.-P., & Liu, Y. (2009). On strategies for imbalanced text classification using SVM: A comparative study. *Decision Support Systems*, *48*(1), 191–201. Retrieved from https://doi.org/10.1016/j.dss.2009.07.011 doi: 10.1016/j.dss.2009.07.011

Webb, G. I., Keogh, E., Miikkulainen, R., Miikkulainen, R., & Sebag, M. (2010). Naïve bayes. In *Encyclopedia of machine learning* (pp. 713–714). Springer US. Retrieved from https://doi.org/10.1007/978-0-387-30164-8$_5$76 doi: 10.1007/978-0-387-30164-8$_5$76

wget: pure python download utility. (n.d.). Retrieved 2022-06-26, from http://bitbucket.org/techtonik/python-wget/

Appendix

A. Tables & Figures

Table 1

journal	retracted	non-retracted	total
arabian journal of geosciences	83	72	155
journal of cellular biochemistry	78	7	85
oncotargets and therapy	13	31	44
rsc advances	72	30	102

2 *group train/test dataset: journals used for training and testing the classifier*

Table 2

journal	retracted	non-retracted	total
journal of fundamental and applied sciences	12	47	59
plos one	60	22	82

2 group external validation dataset: journals used for external validation

Table 3

journal	error	misconduct	non-retracted	total
acs applied materials & interfaces	2	7	31	40
artificial cells nanomedicine and biotechnology	2	12	42	56
biochemical pharmacology	2	2	1	5
biomed research international	5	5	9	19
blood	3	2	14	19
brain research	3	2	15	20
canadian journal of physics	3	3	20	26
cancer gene therapy	2	6	12	20
cancer letters	7	8	13	28
cancer research	6	26	42	74
cell	5	6	40	51
cell cycle	2	5	18	25
cell metabolism	3	4	25	32
construction and building materials	2	6	33	41
embo journal	5	2	2	9
evidence-based complementary and alternative medicine	1	2	4	7
experimental and therapeutic medicine	1	17	56	74
experimental cell research	3	2	20	25
industrial & engineering chemistry research	1	5	9	15
international immunopharmacology	2	2	7	11
journal of biological chemistry	1	8	4	13
journal of bone and mineral research	2	1	15	18
journal of cell science	2	5	16	23
journal of cellular biochemistry	7	70	33	110
journal of cellular physiology	8	18	14	40
journal of controlled release	1	3	7	11
journal of neuroscience	8	4	5	17
journal of the american chemical society	1	2	3	6
lancet	3	2	2	7
life sciences	2	8	22	32
materials science and engineering	2	3	11	16
mathematical problems in engineering	2	3	23	28
medicine	5	2	15	22
naunyn-schmiedebergs archives of pharmacology	2	6	2	10
neurocomputing	2	1	9	12
renewable energy	1	2	11	14
rsc advances	5	20	1	26
scientific world journal	1	3	3	7
thin solid films	2	2	9	13
tumor biology	2	19	74	95

3 group train/test dataset: journals used for training and testing the classifier

Table 4

journals	error	misconduct	non-retracted	total
molecular medicine reports	5	18	92	115
plos one	34	61	157	252

3 group external validation dataset: journals used for external validation

Table 5

error	misconduct
error by journal/publisher	fake peer review
error by third party	false affiliation
error in analyses	false/forged authorship
error in cell lines/tissues	falsification/fabrication of data
error in data	falsification/fabrication of image
error in image	falsification/fabrication of results
error in materials (general)	hoax paper
error in methods	manipulation of images
error in results and/or conclusions	manipulation of results
error in text	misconduct by author
duplicate publication through error by journal/publisher	misconduct by company/institution
	misconduct by third party
	paper mill
	plagiarism of article
	plagiarism of data
	plagiarism of image
	plagiarism of text
	randomly generated content
	sabotage of materials
	salami slicing

List of reasons for retraction for the groups error and misconduct

Table 6

		NB						**SVM**						**LR**					
Group	**Method**	F1		recall		precision		F1		recall		precision		F1		recall		precision	
		t	e	t	e	t	e	t	e	t	e	t	e	t	e	t	e	t	e
non-retracted	un-	.57	0	.40	0	1	.57	.91	.49	.89	.39	.94	.64	.91	.45	.89	.36	.94	.60
retracted	balanced	.86	.68	1	1	.75	.51	.95	.67	.86	.79	.94	.58	.95	.64	.97	.76	.94	.56
weighted avg.		.75	.35	.78	.51	.84	.26	.94	.58	.94	.60	.94	.61	.94	.55	.94	.57	.94	.57

Classifier F1, recall and precision test (t) and external validation (e) scores for the 2 group train/test dataset and external validation dataset, respectively, using different sampling methods

Table 7

		NB						SVM						LR					
Group	**Method**	F1		recall		precision		F1		recall		precision		F1		recall		precision	
		t	e	t	e	t	e	t	e	t	e	t	e	t	e	t	e	t	e
non-retracted		.78	.81	1	1	.64	.68	.83	.78	.84	.81	.82	.76	.85	.84	.92	.94	.79	.76
misconduct	un-	.16	.02	.09	.01	.88	1	.69	.41	.7	.41	.68	.42	.7	.45	.68	.38	.73	.55
error	balanced	0	0	0	0	0	0	.33	.06	.3	.05	.36	.08	.28	0	.17	0	.83	0
weighted avg.		.53	.55	.64	.68	.63	.68	.74	.63	.74	.64	.74	.61	.75	.67	.78	.72	.78	.63
non-retracted	specified							.84	.79	.85	.81	.83	.77	.85	.83	.91	.92	.79	.76
misconduct	class							.68	.4	.68	.41	.68	.4	.69	.45	.66	.39	.72	.53
error	weights							.32	.06	.3	.05	.35	.09	.35	.04	.23	.03	.7	.1
weighted avg.								.74	.63	.74	.64	.74	.61	.75	.67	.77	.71	.76	.64
non-retracted		.75	.72	.7	.63	.8	.83												
misconduct	over-	.55	.49	.67	.7	.47	.38												
error	sampled	.12	.11	.1	.1	.15	.11												
weighted avg.		.62	.6	.62	.59	.64	.66												
non-retracted	sampled	.84	.78	.77	.74	.92	.81	.92	.83	.96	.98	.88	.73	.9	.81	.91	.91	.89	.74
misconduct	by	.5	.21	.52	.16	.48	.3	.65	.38	.59	.27	.74	.64	.61	.33	.62	.27	.6	.43
error	journal	.18	.21	.4	.36	.12	.15	.33	0	.2	0	1	0	.27	0	.2	0	.4	0
weighted avg.		.76	.59	.71	.58	.82	.63	.85	.65	.87	.72	.87	.63	.82	.62	.83	.67	.82	.59

Classifier F1, recall and precision test (t) and external validation (e) scores for the 3 group train/test dataset and external validation dataset, respectively, using different sampling methods

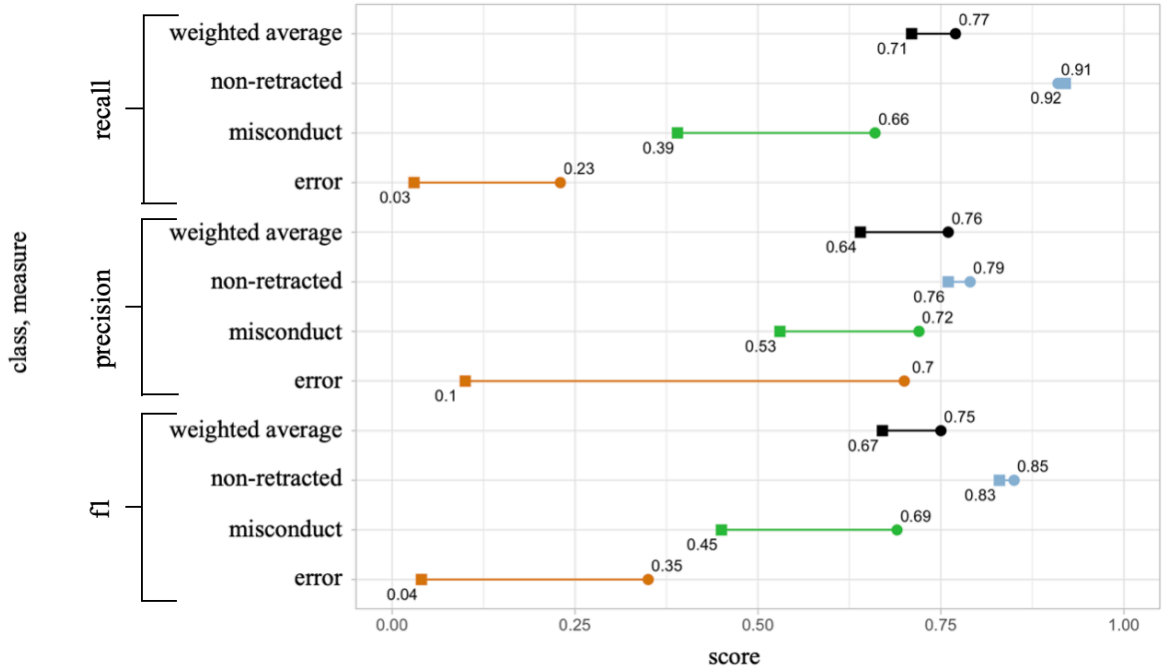
Table 8

Dataset	**Method**	**tf-idf top 30 words**
2 group train/test	unbalanced	'expression', 'si', 'cell', 'group', 'copyright', 'reserve', 'protein', 'tissue', 'protect', 'invasion', 'right', 'cadherin', 'patient', 'article', 'ul', 'tumor', 'scratch', 'migration', 'chamber', 'mrna', 'buffer', 'gene', 'adjacent', 'transfecte', 'assay', 'study', 'decrease', 'staging', 'cancer', 'line'
3 group train/test	unbalanced, specified class weights, oversampled	'film', 'si', 'poly', 'thin', 'laser', 'defect', 'mobility', 'stress', 'carrier', 'density', 'anneal', 'processing', 'grain', 'electron', 'dose', 'roughness', 'residual', 'annealing', 'phonon', 'wavenumber', 'concentration', 'diode', 'temperature', 'nm', 'beam', 'glass', 'dopant', 'quality', 'shift', 'melting'
	sampled by journal	'actuator', 'forest', 'wax', 'regime', 'stiffness', 'strain', 'film', 'corrugate', 'thermal', 'melting', 'composite', 'cnt', 'paraffin', 'melt', 'yarn', 'load', 'heat', 'stress', 'conductivity', 'capillary', 'expansion', 'confine', 'infiltration', 'vertically', 'vertical', 'pressure', 'expand', 'nanocomposite', 'shape', 'compress'

Top indicative words for the 2 groups, 3 groups (and different sampling methods). The tf-idf top 30 words were the same for all classifiers, but different for 2 and 3 group train/test datasets (which were used for fitting the classifiers) and some of the sampling strategies. For 3 groups, tf-idf top 30 words for the sampling methods unbalanced, specified class weights, oversampled were the same.

Figure 7

Figure 7
LR with specified class weights, F1, recall and precision scores for the classes error, misconduct, non-retracted and weighted average, using the 3 group train/test and external validation datasets



Points represent test scores using the 3 group train/test dataset, squares represent scores for external validation using the 3 group external validation dataset. Black represents the weighted average of the classes of the specific measure, light blue represents the non-retracted, green the misconduct and orange the error class.

B. Notebooks

Notebooks for data acquisition, pre-processing and analysis for the 2 group (non-retracted, retracted) and the 3 group (non-retracted, misconduct, error) train/test and external validation datasets can be found in my Github repository:
`https://github.com/Arl-cloud/Thesis`