# Can linguistic features unmask potential fraudulent research?

Distinguishing retracted and non-retracted papers
using an NLP classifier based on text and linguistic features

**Master thesis**

|  |  |
|---|---|
| Author: | Eveline Liselotte Schmidt, MA |
| Programme: | MSc Applied Data Science |
| University: | Utrecht University |
| First supervisor: | dr. Javier Garcia Bernardo |
| Second supervisor: | dr. Ayoub Bagheri |
| Date: | July 1st, 2022 |

# Abstract

Researchers experience a lot pressure to get published and cited, as their careers often depend on it. This pressure can result in various forms of misconduct. Fraud in academic research is an important problem that should be tackled. Text classification is one way how fraudulent papers can be detected. This project shows that a Logistic Regression classifier can distinguish retracted papers from non-retracted based on texts. This is only possible for papers within the same topic and journal as the classifier was trained on. The results are not generalisable to more general papers or other topics. Literature suggests there are linguistic markers for deceptive language. In this project the features quantity of lexicon, readability, complexity, lexical diversity and number of references are analysed. The quantity of lexicon, complexity and lexical diversity showed significant differences between retracted and non-retracted papers. Including these five linguistic features did, however, not improve the performance of the classification model.

## Keywords

Fraudulent research, Deceptive Language, Text classification, Stacking classifiers

# 1. Introduction

## 1.1 Motivation and context

The pressure to publish and be cited has been high for researchers for long time. A researcher's success is often dependent or pictured by the amount one has been cited throughout their career. To get citations, the paper of the research needs to be published by a journal. Not every research outcome is as easily being published as others. Research with significant results is more likely to be published than research with non-significant results (Head, Holman, Lanfear, Kahn, & Jennions, 2015). Especially prestigious journals tend to publish research that show significant results more than research with insignificant results. Whether you get funding for your research or a job in academia, is also dependent on the number of publications and the journal your research is being published in, as those are often seen as a measurement to determine one's academic competence. Hence, it is obvious that most researchers prefer to be published in a prestigious journal. These researchers therefore hope to retrieve significant results. When statistical tests give the researcher a p-value of 0.051, after a lengthy period of working hard and putting together a good research, it can be very unmotivating, as one knows their chance of being published has decreased. Combined with the pressure the researcher feels, from itself or from the institution that provides funds for the research, one might be tempted to include or exclude an extra variable in order to get a p-value like 0.049. This phenomenon is also known as p-hacking. It is one example of how fraud can be committed in research, but there are many other ways. The most obvious way of fraudulent research that most people think of is plagiarism. Plagiarism is relatively easy to check. There are many algorithms that can see to what extent a paper has similarities with text on the internet or with other research papers. Other kinds of committing fraud, such as p-hacking, are less obvious at the first sight. These kinds of committing fraud are therefore harder to detect but are not less important to be discovered. This project will therefore focus on trying to build a classifier that can distinguish retracted academic papers from non-retracted academic papers. The associated research questions of this project will be given and elaborated at the end of this chapter.

## 1.2 Literature overview

### 1.2.1 Different kinds of fraudulent research

According to the U.S. Office Science and Technology Policy, and many other research funding agencies scientific, there are three fundamentals of misconduct in research: fabrication, falsification and plagiarism (2009). With fabrication it is meant when data or the results are made up, whereas with falsification it is meant that research material, equipment or processes are manipulated, by changing or disregarding data or results so that is it not complete and accurate anymore. Plagiarism on the other hand is seen as the act of copying someone else's results, processes, words or ideas without giving them credit for it. Besides these three fundamentals, the U.S. Office Science and Technology Policy states that the misconduct has to be executed intentionally and knowingly. This addresses an important remark that has to be made. It is important to differentiate between an intentional error and an unintentional mistake. If someone unknowingly or by mistake does not include something in their research or forgets to check something, it might be unfair to immediately see that person as a fraud. However, even if someone manipulates the data intentional, they could always argue it was a mistake. The line between an intentional error or an unintentional mistake is very thin and probably also subjective.

Another term for falsification, which is used in the academic world often and that was quickly mentioned before as well, is p-hacking. P-hacking is the phenomenon when researchers keep on collecting or select specific data in order to change statistical insignificant results to significant results (Head, Holman, Lanfear, Kahn, & Jennions, 2015). However, p-hacking does not necessarily alter the scientific agreement from meta-analyses. In order to prevent p-hacking and other forms of fraud, researchers are more often asked to be open and clear about their data, approaches and analyses. However, it is not only the responsibility of the researchers themselves according to many, but also of the journals that publish the papers. If journals would require publication of sample size rules, measurements and alterations fraud could be even more countered (Simmons, Nelson, & Simonsohn, 2013).

### 1.2.2 Quantity of fraudulent research

Every once in a while the news comes out that a (well-known) researcher has been accused of fraud. This is often devastating for their careers and personal life, as their other research is often also questioned. In most cases, the misconduct has been committed intentionally or inadvertedly and therefore the researchers themselves will not admit it easily and voluntarily. Hence, it is hard to say how often exactly fraud has been committed in research. Also colleagues of fraudulent researchers are not very likely to address their suspicion on their fellows to the appointed committees, due to cronyism. It is thus likely that there is still a lot of not yet discovered fraudulent research going around. This makes it hard assess how much misconduct is happening in academia. Pupovac and Fanelli performed a meta-analysis on surveys that asked researchers about misconduct in their field (2015). They found that the rate of researchers that report knowing about their colleagues committing plagiarism is higher than that of fabrication and falsification. When it comes to researchers themselves, the rate of admitting misconduct, regardless of the type, has declined over time. This could either mean that researchers are committing less often fraud or that they are nowadays less honest about it.

### 1.2.3 Detecting fraudulent research

The quality of science decreases due to the pressure to publish (Sarewitz, 2016). It is obviously important that the quality of science is kept high. Therefore fraudulent research needs to be discovered and penalized. There are multiple approaches that try to find fraudulent research. One possibility is to check whether the results and outcomes of the research are in fact possible. Schumm, Crawford and Lockett suggest the use of statistics from binary variables to detect anomalies in the data and with that potential fraudulent research (2019). A wide variety of tests have been applied to detect potential fraudulent research. An example of such a test is the GRIM Test (Granularity-Related Inconsistency of Means) that tests whether the means of the data are consistent with the sample size and amount of instances in the data (Brown & Heathers, 2017). The study of Brown and Heathers (2017) shows that 189 papers out of a total of 260 (73%), show some level of problems according to the GRIM Test, which is a worryingly high amount. The level of these problems varies from minor to substantial.

### 1.2.4 Deceptive language and linguistic markers

The aforementioned examples that try to discover fraudulent research focus only on the numerical data used in the research itself. There are also other possibilities that can potentially discover fraudulent research that do not use numerical data. Another way to detect potential fraudulent research, is to look at the text of the academic paper. An example of such is the SCORE program (Alipourfard, et al.). This program builds and verifies algorithms that check claims in research papers, by using amongst others machine learning. Language use tends to be a good indicator of deception (Bachenko, Fitzpatrick, & Schonwetter, 2008). The detection of deceptive language in written and spoken language is applied in many areas to catch deceivers. Examples are the detection of fraud in annual reports (Goel & Gangolly, 2012) or to decide whether an interrogator is lying (Porter & Yuille, 1996).

Research shows that there are several linguistic features that predict deception in a text (Burgoon, Blair, Qin, & Nunamaker, 2003). The linguistic cues that are ought to be markers of deceptive language are however not consistent throughout different research (Levitan, et al., 2015). Combining such linguistic features together improves the capability of identifying deception in texts (Burgoon, Blair, Qin, & Nunamaker, 2003). An aspect that influences the results of those identification techniques, that has to be mentioned, is the topic of the texts. Newman, Pennebaker, Berry and Richards used a computer-based text analysis program to identify deceptive texts from truthful texts (2003). When the topic of the texts stayed the same the accuracy was higher (67%) than when the analysis was performed on multiple topics (61%).

## 1.3 Classifying fraudulent research from non-fraudulent research

In this study the focus will lay on five linguistic features of which research has shown they are indicative of deceptive language. These five features are: quantity of lexicon, readability, complexity, lexical diversity and the number of references. Below an overview will be provided that gives a broader explanation for each feature and how it relates to deceptive language.

### 1. Quantity of lexicons

A lexicon, in the case of academic papers, refers to the words used in that paper. Research shows that deceptive language has a higher amount of lexicons, such as verbs, nouns and modifiers (Mbaziira & Jones, 2016). A reason for this higher amount could be that the fraudster wants to cover up its fraud by using a lot of words. Truthful language is often more to the point and on a more detailed level (Markowitz & Hancock, 2016).

### 2. Readability

Deceptive language tends to have a higher degree of linguistic obfuscation than truthful language (Markowitz & Hancock, 2016). Linguistic obfuscation is the use of unclear language to confuse the reader and is characterized by lower levels of readability. It can therefore be stated that deceptive language tends to have a lower readability. An explanation for this could be that the fraudsters want to confuse the reader intentionally to make them not question (the quality of) the research.

### 3. Complexity

Deceptive language uses fewer complex words than natural language (Newman, Pennebaker, Berry, & Richards, 2003) (Zhou & Sung, 2008). Deception is a complex and cognitive consuming task for humans, as deception uses more cognitive processing than telling the truth does. Deceivers have to keep up normal communication, while they are also thinking about the lie(s) they are trying to hide. This increase in cognitive processing will cause the deceiver to use easier language. This is also supported by the fact that for deceptive language the average word length also tends to be a predictor (Afroz, Brennan, & Greenstadt, 2012). Shorter words are often less complex than longer words and thus more often used in deceptive language (Newman, Pennebaker, Berry, & Richards, 2003).

### 4. Lexical diversity

Research shows that deceptive language has less unique words in its lexicon than truthful language does (Mbaziira & Jones, 2016). A lower ratio of unique words refers to a lower lexical diversity. The lexical diversity could point towards the freedom to choose your own words authors feel when writing. Deceivers do not want to be unmasked, so they will try to stay to a strict script and could potentially feel less free to use a large variety of words.

### 5. Number of references

One of the most consistent markers of deceptive linguistic style is the use of pronouns. There is, however, no consensus on the direction of this kind of marker. Some state that deceptive language uses less self and other references (Newman, Pennebaker, Berry, & Richards, 2003). Others also state that deceptive language tends to use less first-person singular pronouns, such as ‘I’, ‘me’ and ‘my’ than non-deceptive language (Hancock, Curry, Goorha, & Woodworth, 2007). However, they also state that the usage of second- and third-person singular pronouns, such as ‘you’ and ‘she’, is more present in deceptive language. A pronoun is used to refer to either yourself or to someone else. As in academic papers it is unconventional to use pronouns, it might be irrelevant to investigate the usage of pronouns. However, there is another way to refer to others in academic papers, which is by referencing other papers. This could be seen as a form of second-person singular pronouns. The number of papers that is referenced to could therefore be a predictor for fraudulent research. Fraudulent papers tend to reference more often than non-fraudulent papers (Markowitz & Hancock, 2016). A possible explanation for this could be that the fraudulent researchers are trying to cover up their fraud by citing more other researchers to support their claims, results and approaches.

As can be concluded based on the aforementioned literature, there is enough reason to investigate whether it is possible to create a paper classifier while using natural language processing (NLP). The focus of this study will lay on solely academic papers. The purpose of this current study is to build a classifier that tries to distinguish fraudulent papers from non-fraudulent papers. The accompanying research question [RQ1] is therefore:

*'Can a classifier distinguish fraudulent research papers from non-fraudulent research papers?'*

The input for this classifier will first consist out of only text from the academic papers. After that, the textual input will be accompanied by linguistic features based on the texts. This will answer the second research question [RQ2]:

*'Will including linguistic features result in better outcomes for classifying fraudulent research papers from non-fraudulent research papers?'*

# 2. Data

## 2.1 Data selection and data scraping

When papers are found to be fraudulent, the journal that has published the paper, will retract the paper. The journal will publish a formal withdrawal of the published paper. The paper itself will still be available on the internet, but in most cases the paper has a special notice or watermark that informs the reader of its retraction.

To be able to see whether fraudulent papers can be distinguished from non-fraudulent papers, based on written text and linguistic cues, a dataset is needed that includes both retracted and non-retracted academic papers. The final database shall consist out of two different databases that are concatenated. One database includes all retracted papers, whereas the other database only includes non-retracted papers. For the retracted academic papers, a special dataset is used from The Centre For Scientific Integrity, the parent non-profit organization of Retraction Watch. This dataset includes many retracted papers, which are retracted for several reasons, like plagiarism of text or falsification of the results, but also due to errors in the data. As it also includes erroneous research, it is important to note that retracted papers are not always fraudulent. For the non-retracted papers, a dataset was made that consists out of papers that were very similar to the retracted papers, based on their topic and journal, but which were not retracted. The procedure for finding these papers and making the datasets as similar as possible shall now be explained.

At a fist attempt to make comparable datasets, one topic was chosen to select both retracted and non-retracted papers about. This topic was 'environmental studies'. A first exploration of the data showed that the results were very topic-specific and not generalisable to other topics. Therefore, I decided to include more than one topic. To let the topics be as comparable as possible, I chose to select not topics, but journals. Journals tend to be built around a specific topic or research field. Another advantage of including journals is that the outline of the articles is also highly comparable. To make results more generalisable, multiple journals were selected. Results would this way probably be less topic specific.

To decide on which journals shall be selected, an exploration of the available journals was performed on the database of retracted papers. This exploration was done on the retracted papers, as it would be easier to retrieve more non-retracted papers than retracted papers. The journals are selected based on their field of research, combined with the amount of papers available. An overview of the journals and the number of papers, that is included in the database, can be found in table 1.

The database of the retracted papers did not yet include the texts of the papers. Therefore the papers needed to be scraped from the internet, using the DOI of the papers. Also the non-retracted papers are scraped from the internet, using the python library Beautiful soup. PDFs from both kinds of papers were automatically scraped from the internet and downloaded into a private Google Drive folder.

*Table 1 Overview of final datasets*

| Journal name | Usage | Number of both retracted and non-retracted papers |
|---|---|---|
| Arabian Journal of Geosciences | Internal | 47 |
| International journal of Electrical Engineering & Education | Internal | 32 |
| Journal of Cellular Biochemistry | Internal | 78 |
| OncoTargets and Therapy | Internal | 13 |
| Journal of Fundamental and Applied Sciences | Internal | 12 |
| RSC Advances | External | 72 |
| PLoS One | External | 60 |

## 2.2 Data pre-processing

Once the PDFs of the papers were collected, the textual data of those papers could be retrieved. With the use of the module fitz in Python, the text was read from the PDF and put into a data frame that consists out of a paper ID, the content and whether it is a retracted paper or not.

The content was then analysed for missing values. In case of missing values in the content column, the whole instance was removed. These missing values were probably due to the fact that not all papers were scraped or transformed to text correctly.

While analysing the data, I discovered that some retracted papers included a Retraction Notice, which is as mentioned before, a common way to inform the readers of the paper of the retraction. The whole notice and the words 'Retraction, retraction, Retracted, retracted, retract, Retract, withdrawn, Withdraw' were removed from the texts. These words were also removed from the non-retracted texts so that they would not be seen as indicative for non-retracted papers by the classifier. Besides that, all numbers, so also including years, were removed from the texts.

The texts still included a lot of noise such as other information sheets from the journals. Therefore, I decided to split the content of the texts on the sections of the papers. The content now consisted out of the text starting at the Introduction of the paper up till the Reference list of the paper. If a paper did not include the chapters 'Introduction', 'Discussion' or 'Conclusion' and 'References', the paper was not considered to be an academic paper and was therefore removed. At last duplicates were checked and removed, if present.

The initial scraping was not successful in all cases. In some cases, the paper was not available and was therefore skipped by the algorithm. Combined with the removal of papers due to various aforementioned reasons, the dataset was disbalanced. As it is important that a specific journal, and thus topic, is not much more included in either one of the groups, the databases were balanced by under sampling on journal. In the final dataset, seven journals are left that include as many retracted as non-retracted papers. In order to test whether the fitted classifiers are generalisable. The fitted classifiers will be tested on journals it has never seen before. To do this, a distinction is made between an internal and external dataset. The internal dataset consists out of five journals that will be used to train and test the classifier. The external dataset consists out of two journals, that will be used to test the fitted classifier its generalizability. These two journals are specifically selected, due to the fact that their topics are of a general and diverse nature. The journal RSC Advances publishes research on all aspects of chemical science and PloS One on all aspects of biological science. An overview of the final datasets can be found in table 1. These final datasets are the same as the dataset of Franssen (2022).

In order to not let the classifiers classify based on noise from the journal formats, all proper nouns, so author names, but also journal names were removed. In order to make the texts best analysable and interpretable for classifiers stop words and punctuation marks are removed. Besides that all words are lowercased and lemmatized.

## 2.3 Ethical and legal considerations of the data

All data from both retracted and non-retracted papers is accessible through my institution. Therefore there are no legal or privacy-related issues attached to this data. It is, however, important to note that it can do a lot of harm when people are accused of being fraudulent in their research. When researchers are accused of being fraudulent in a certain research, their whole career is often questioned (Smith, 2005). This can happen on individual level, but also on institutional or even research fields level (National Academy of Sciences, National Academy of Engineering (US) and Institute of Medicine (US) Committee on Science, Engineering, and Public Policy., 2009). Therefore, I decided to not perform predictive analyses on data that questions the truthfulness of academic papers which have not (yet) been accused of misconduct.

Even though all information in the dataset of The Retraction Watch Database is available of the associated website, the full dataset itself is not publicly downloadable. For this project, access to this dataset was granted. Before the start of this project an agreement was signed that the dataset shall only be used for this research and shall not be published online. The dataset has only been stored on a private Google Drive and was not published on publicly accessible sites such as GitHub.

Regarding ethical considerations of the data and analyses, I decided to not use personal information on author level as predictors for fraudulent papers. Even though this has been done by others in the past, for example by looking at the country or language of the author (Stretton, et al., 2012).

# 3. Methods

## 3.1 Classification models

A classification model is a type of model that reads input and classifies it into distinct categories. The number of categories can differ. There is binary classification, in which the model classifies the output into one of two categories. Another possibility is multi-class classification where the model classifies the output into one of more than two categories. In this project there are only two categories, i.e. retracted and non-retracted papers, which makes it binary classification. Classification can be done via either supervised or unsupervised learning. This means that the categories can be determined beforehand, but also later on, by the classifier itself. Given the fact that the categories are determined beforehand, this project is an example of supervised learning. This project concerns the classification of texts of academic papers, which makes it text classification. This text classification is done by Natural Language Processing (NLP). There are many models that can perform this type of binary text classification. Examples are classification models such as Naïve Bayes, Random Forest, Support Vector Machine, Logistic Regression and BERT.

### 3.1.1 Logistic Regression

Of these classifiers, Logistic Regression is a very popular classifier, which is often used for classification and regression. The underlying technique is the same as for Linear Regression. Logistic Regression is often used for binary classification. It is one of the classic types of classifiers, as it has been used for over a long time and its steps are all explainable, unlike newer types of classifiers such as BERT. Classifiers like BERT are forms of deep learning language models that have been pre-trained on substantial amounts of texts. These algorithms are very complex and hard to grasp by humans.

In this project, Logistic Regression will be used to answer both research questions. For the second research question [RQ2], text classification is combined with classification based on numerical data. Logistic Regression can be used for classification of both textual and numerical data. Most classifiers work directly on numerical data when it comes to classification of that data. However, when it comes to text classification, the data consists, obviously, out of texts. These texts have to be converted to interpretable numbers for the classifiers. This can be done with either Count Vectorizer or TF-IDF vectorizer. For Logistic Regression, the text should also be converted to numerical data. For this project I use Count Vectorizer. Count Vectorizer transforms the text to a sparse matrix. In this matrix, the columns represent all unique words that are present in the texts and the rows represent the word count of those words per document or sentence.

Hyperparameter tuning can optimize the performance of classifiers (Elgeldawi, Sayed, Galal, & Zaki, 2021). Due to the scope of this project, finding the optimal settings for the hyperparameters manually was not possible. Elgeldawi, Sayed, Galal and Zaki (2021) showed that in order to retrieve the best performance of Logistic Regression for sentiment analysis, the following settings are suggested: *solver = 'saga', penalty = 'l2', multi_class = 'auto', max_iter = 100, C = 10*. As sentiment analysis also uses text classification (Gupte, Joshi, Gadgul, Kadam, & Gupte, 2014), the same settings can be used in the Logistic Regression analyses of this project, to distinguish retracted papers from non-retracted papers. The suggested settings are very similar to the default settings of the hyperparameters. Only the solver, the algorithm used in the optimization problem, and C, the inverse of regularization strength, are different from default. A higher value is now given to C, which means that more weight is given to the training data. The *saga* solver is, according to documentation, mostly useful for large amounts of texts. As this is not the case for the datasets in this project, the solver *liblinear* is used, as suggested by the documentation to do so for smaller datasets.

### 3.1.2 Stacking classifiers

For this project Logistic Regression was performed several times on the data. For the first research question [RQ1] ‘Can a classifier distinguish fraudulent research papers from non-fraudulent research papers?’ the classifier was only applied on textual input. For the second research question [RQ2], ‘Will including linguistic features result in better outcomes for classifying fraudulent research papers from non-fraudulent research papers?', on the other hand, two classifiers were used at the same time. This phenomenon is called 'stacking classifiers'. This means that multiple classifiers are used together to train a meta-classifier. This meta-classifier's input consists of the output of the multiple classifiers and decides based on this input to which class the instance will be mapped. The idea behind stacking classifiers is that one classifier knows less than multiple. Stacking classifiers is often done to improve model predictions. In this project the stacking classifier will consist out of two Logistic Regression classifiers. One is the aforementioned classifier, that uses the texts of the papers as input. The other one is a classifier that uses linguistic features from those texts as input. How these linguistic features are operationalized is explained in the next section 3.2 Linguistic feature analysis.

## 3.2 Linguistic feature analysis

Previous research has shown that there are several linguistic features that have proven to be able to predict deceptive language. As stated before, there are five linguistic features that will be included in the input of the classifier. These five features are: quantity of lexicons, readability, complexity, lexical diversity and number of references. Including these features into the classifier and comparing the results to the results from the classifier that uses only text as input will answer the second research question [RQ2] 'Will including linguistic features result in better outcomes for classifying fraudulent research papers from non-fraudulent research papers?'. For every feature, a short operationalization is provided below, that shows how the feature will be measured and what outcome it expected based on the previously mentioned literature. The features will be retrieved via the use of the python libraries readability and textstat.

### 1. Quantity of lexicon

Deceptive language tends to have a higher quantity of lexicon than non-deceptive language, as mentioned before. The level of abstractness will be measured by looking at the word count. Obviously, the length of papers differs per author, journal and or topic. This should be overcome by the fact that the datasets are balanced by journal. It is expected that the retracted papers have a higher quantity of lexicon, i.e. use more words, than the non-retracted papers.

### 2. Flesch Reading Ease

Research showed that deceptive language has a lower readability than truthful language. There are several measurements for readability. For this study, the Flesch Reading Ease is used, as it is an old and accurate measurement for readability. It uses two features that determine the final score: the average number of words in a sentence and the average number of syllables per word. The final score can range between 0 and 100 (Eleyan, Othman, & Eleyan, 2020). A score between 90 and 100 means that the text is very easy to read, usually doable for a fifth grader. A score between 0 and 30 on the other hand, points towards the fact that the text is hard to read and that it is best understood by a college graduate or professional. As all papers are academic papers, the scores are expected to be low. However, for the retracted papers, a lower score is expected than for the non-retracted papers.

### 3. Complexity

To measure the complexity, the number of complex words will be retrieved via the readability library. The total amount of complex words in the paper is divided by the total amount of words in that paper, in order to normalise the data. A word is considered to be complex if that word has three or more syllables. As stated before, deceptive language is often less complex and has a lower average word length than truthful language. Therefore, the amount of complex words is considered as a linguistic feature that could indicate misconduct.

**4. Lexical Diversity**

The Type Token Ratio (TTR) is an equation that measures the lexical diversity. It divides the total number of unique words (types) by the total number of words (tokens). Deceptive messages have shown to have fewer ratio of unique words and thus a lower lexical diversity. A low lexical diversity is expressed in a low TTR, which means that the retracted papers are expected to have lower lexical diversity.

**5. Number of references**

Research showed that fraudulent research tends to reference to other papers more often. To measure the number of references, the length of the reference list will be investigated. This is not a fool proof method, as the title length of the paper and the journal will also influence this measurement. However, the amount of papers that are included in this project will try to overcome that.

## 3.3 Metrics

To be able to compare the performance of the different classifiers, several metrics will be analysed. These metrics are accuracy, recall, precision and F1-scores. They are based on the output from the confusion matrix that provides the amount of True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN), as shown in table 2. How these metrics are calculated can be found in table 3. Each metric has its own advantages and disadvantages. Accuracy is an often used and straightforward metric, that tells how often the model has been correct. However, in case of unbalanced datasets, accuracy may not be reliable. If one group consists out of 95 instances and the other group of 5 instances, the accuracy will still be 95% when the classifier classifies all instances to the first group. This result might seem good, but if it was the intention to find the instances in the small group, which is often the case when detecting fraud, the model did not perform good at all. In this project, the groups are balanced. Therefore, this will not be a problem. However, it is still good to look at other metrics that deal better with such complications.

Examples of such metrics are precision and recall (Davis & Goadrich, 2006). Precision focusses on the performance of the model, by looking at the amount of predicted positive instances which are correctly classified as positive. Whereas recall, also known as sensitivity, measures the performance by looking at the amount of truly positive instances that are correctly predicted positive.

At last, the F1-score is a metric that uses both the precision and recall to measure the performance (Powers, 2020). It takes the harmonic mean of both metrics.

As all metrics are informative in their own way, all will be provided when discussing the results. The focus will, however, lay on the F1-scores, as these are a mix of recall and precision and give a good indication of the performance in most cases. This is because the F1-score takes on one hand, into account that all papers that should be retracted, are retracted (measured by recall). While on the other hand, it also takes into consideration that it does not accuse papers of fraud which are in fact not fraudulent (measured by precision). This is very important as accusing one of fraud is a very big and impactful matter.

*Table 2 A confusion matrix*

| | Predicted as non-retracted papers | Predicted as retracted papers |
|---|---|---|
| Actual non-retracted papers | TN | FP |
| Actual retracted papers | FN | TP |

*Table 3 Overview of the performance measures and their associated calculations*

| Performance measure | Calculation |
|---|---|
| Accuracy | $\frac{TP + TN}{TP + TN + FP + FN}$ |
| Precision | $\frac{TP}{TP + FP}$ |
| Recall | $\frac{TP}{TP + FN}$ |
| F1-score | $\frac{2 \times Precision \times Recall}{Precision + Recall}$ |

# 4. Results & Discussion

The Python Notebooks of the aforementioned classifiers and linguistic analyses, along with the results that shall now be discussed, can be found on my GitHub page, as stated in Appendix 1.

## 4.1 Classification based on text only

At first, Logistic Regression was performed on the texts of the papers only. The performance of the first classifier is shown in figure 1, for both internal and external datasets. As mentioned before, first the model will be trained and tested on five journals. This can be seen as the internal dataset. However, these results may not always be as reliable as one would expect. Therefore, the fitted classification model is also applied on the external dataset, which includes papers from two other journals it has never seen before.

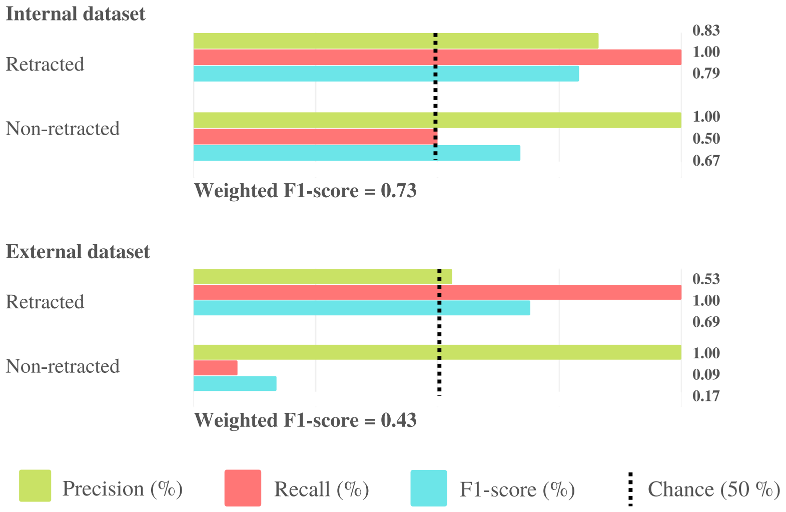


*Figure 1 Performance of Logistic Regression with textual input*

Figure 1 shows that the model performs good on the internal dataset (Weighted F1-score of 0.73). The performance on the external dataset is bad (F1-score of 0.43). This means that the model is not performing better than chance for detecting papers from journals it has never seen before. The recall for the retracted papers is for both internal and external datasets higher than the precision. This means that the classifier tends to classify most instances as retracted. The classifier is thus very good at finding all papers that should be retracted, but it also classifies many non-retracted papers as retracted and with that 'accuses' them of potential fraud. This can also be seen in the lower precision for the retracted papers.

The first research question [RQ1] 'Can a classifier distinguish fraudulent research papers from non-fraudulent research papers?' can thus be answered with the following answer: it is possible to make a classifier that performs better than chance, only if the topics of the papers stay consistent. The classifier is not generalisable. It is not possible to make a classifier solely based on text that performs better than chance, for papers coming from other journals or with other topics. This is in line with previous research that also showed that the performance of a classification model was better when the topics of the text stayed the same (Newman, Pennebaker, Berry, & Richards, 2003).

Even though the results are not generalisable, it could be argued that the academic journals themselves could use such classifiers in order to discover fraudulent papers, as the topic within a journal often stays the same.

## 4.2 Linguistic feature analysis

To see whether the performance of the classification model improves when linguistic features are included in the model, it is important to first look at the features themselves and see whether they are in line with the expectations. The calculations for the features were performed on the internal dataset, together with the external dataset, which makes it a total of papers from seven journals. For each feature, the averages and distribution of the data were investigated. Besides that, the differences were statistically analysed. The output of these analyses and statistical tests can be found in Appendix 2. At the end of this section an overview of the main results is given.

### 1. Quantity of lexicons

The level of abstractness is measured by calculating the number of words for both types of papers. An investigation of the number of words used in the papers, shows that on average retracted papers contain 4698 words (sd = 2192.3), whereas non-retracted papers only include 4482 words (sd = 2365.5). A Mann Whitney U test shows that the difference in quantity of lexicons is significant (p = 0.04623, Cliff's Delta = - 0.09193071). It can thus be stated that the retracted papers use more words on average than the non-retracted papers. A visualisation of these results can be found in figure 2. The distribution of the papers is comparable between the two categories. As can be seen, the retracted papers include a very long paper, which could be an outlier. This outlier could be a possible explanation for the higher average usage of words of the retracted papers. The results match the expectations, that deceptive language uses more words on average and is thus less concise. An explanation for this could be that liars are less concise in their wordings, because of the effort they put in covering up their fraud.

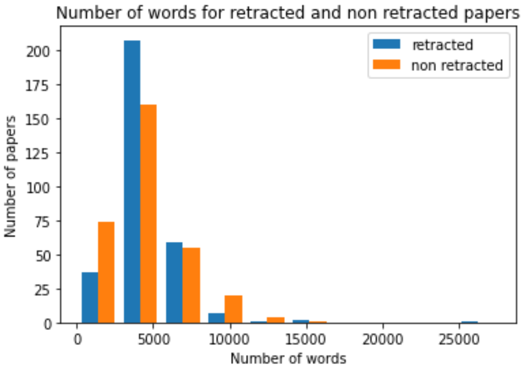


*Figure 2 Distribution of quantity of lexicon*

### 2. Readability

To measure the readability, the Flesch Reading Ease index was chosen as a measurement. Results show that the average Flesch Reading Ease for retracted papers is 42 (sd = 7.3) and for the non-retracted 41 (sd = 10.2). The Flesch Reading Ease ranges from 0 to 100. As stated before, a low score is expected for both papers given the fact that a low score means that it is best understood by a university graduate, which is the reading audience for academic papers. A Mann Whitney U test shows that the difference in the Flesch Reading Ease is not significant (p = 0.2773, Cliff's Delta = - 0.05010345). This means that there are no differences in readability between retracted and non-retracted papers. This is not in line with the expectations, as research has shown that a lower readability is often a marker of linguistic obfuscation and deceptive language. The distribution, shown in figure 3, also shows that the non-retracted papers tend to have an equal readability as retracted papers have. There are some papers with a very low readability, which means that it is very hard to read, but there are also some papers that are easier to read given their higher readability. A possible explanation for these unexpected results, could be that a higher readability could mean less detailed and more general language. Deceptive language tends to be less detailed than non-deceptive language. This could be due to the fact that the author simply does not know the details because the results are (partly) non-existent. Another explanation could be that the Flesch Reading Ease might not be the right metric to measure the readability of academic papers. Academic papers tend to be written specially for a highly educated audience and not for fifth graders for example. The Flesch Reading Ease might be a too general metric that measures readability.

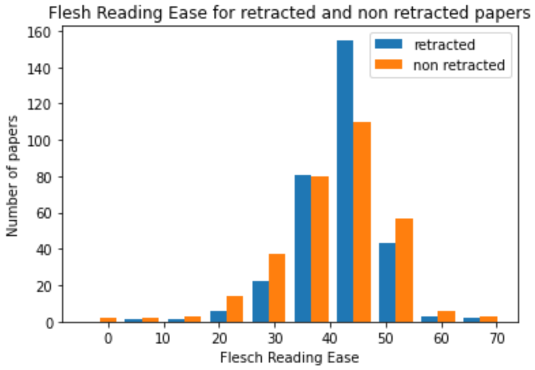


*Figure 3 Distribution of Flesch Reading Ease*

### 3. Complexity

Complexity tends to be a marker of deceptive language. Therefore, the usage of complex words in both types of papers is investigated. The percentage of complex words, normalized by the total number of words, is for retracted papers 21% (sd = 2.4) and for the non-retracted papers 22% (sd = 3.7). A distribution of the percentage of complex words for retracted and non-retracted papers can be seen in figure 4. It shows that the percentage of the complex words of the non-retracted papers is more evenly distributed than that of the retracted ones. A Mann Whitney U test shows that the difference in the use of complex words is significant (p = 0.0003835, Cliff's Delta = 0.1637693). Non-retracted papers use thus on average more complex words than retracted papers. This is in line with other research which showed that deceptive language uses fewer complex words than non-deceptive language. An explanation for this higher percentage could be that committing fraud it a complex and cognitive consuming task for humans, which leaves less cognitive processing left for other tasks such as using complex language. In the readability library a word is considered to be complex when it has more than three syllables. Complex words are in this case thus very comparable to long words. This is also in line with previous research that showed that deceptive language uses more shorter words than non-deceptive language.

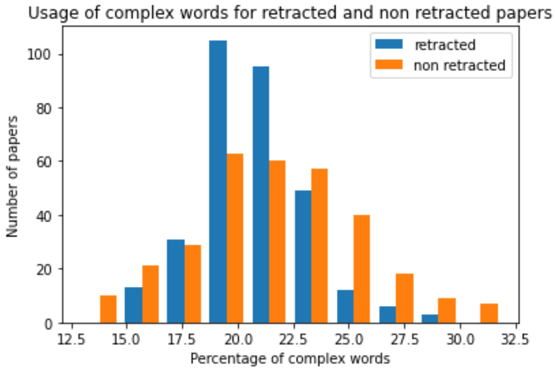


*Figure 4 Distribution of usage of complex words*

### 4. Lexical Diversity

Lexical Diversity is in this study measured by the Type Token Ration (TTR). Research showed that deceptive messages tend to use unique words less often and similar words more often. These findings are in line with the findings of this study that compared the TTR between the two types of papers. The results of this study show that the average TTR for retracted papers is 0.25 (sd = 0.04), which is lower than that for the non-retracted papers, which is 0.27 (sd = 0.09). A distribution of the TTR for both categories can be seen in figure 5. A high TTR means high lexical diversity. The figure shows that most instances have a TTR between 0.2 and 0.3, where the non-retracted papers have more instances with a higher TTR than the retracted papers. A Mann Whitney U test shows that the difference in TTR is significant (p = 0.004551, Cliff's Delta = 0.1308471). The non-retracted papers thus use on average a more diverse lexicon than the retracted papers. These results are in line with the expectations, as deceptive language often has a lower lexical diversity than truthful language. A reason for this could be that the authors of non-fraudulent papers feel more freedom in choosing their words, whereas authors that committed fraud feel less free, as they will try to not be caught.

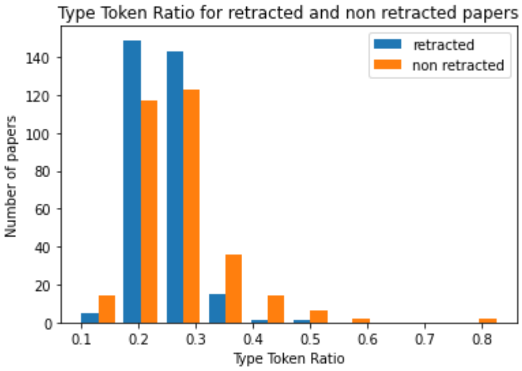


*Figure 5 Distribution of lexical diversity*

## 5. Number of References

As stated before, the use of referrals, often in the form of pronouns, is indicative of deception. Due to the fact that in academic papers the usage of pronouns is unconventional, I decided to use the number of references as an indication of referrals. More specific, the length of the References-section is investigated. Results show that the average length of the References-section is 7235 words (sd = 4875.6) long for the retracted papers and 7419 (sd = 6966.1) for the non-retracted papers. A Mann Whitney U test shows that the difference in the length of the References-section is not significant (p = 0.1799, Cliff's Delta = - 0.06185849). The distribution of the length of the References-section can be found in figure 6 and also shows a very similar distribution for both types of papers. The non-retracted papers thus have, on average, an equally long References-section as retracted papers. This is not in line with the aforementioned research that showed that fraudulent papers reference more often than non-fraudulent papers. It is also not in line with the research that showed that deceptive language uses less other-references than truthful language. It could be that either is still the case for this data, but that even after the pre-processing there is still too much noise in the data. An example of such noise could be that a journal ends the article with a special notice or that the Appendix is also included in the References-section. However, it is unlikely that journal-specific designs influence the results due to the fact that the datasets are balanced on journals and should therefore not play a role.

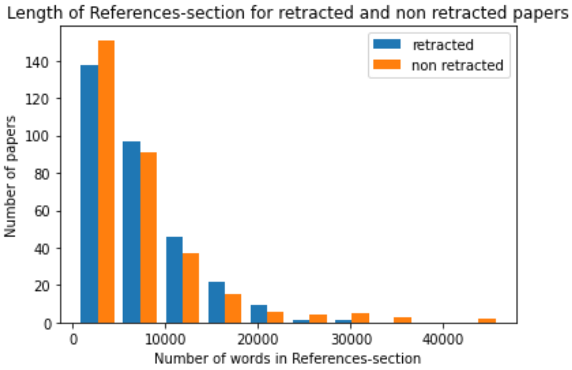


*Figure 6 Distribution of number of references*

**Overview of results**

Table 4 summarises the retrieved results from analyses on the five features.

*Table 4 Overview of the results from analyses on the five linguistic features*

| Feature | Average value for retracted papers | Average value for non-retracted papers | Significantly different? |
|---|---|---|---|
| 1. Quantity of lexicon | 4698 words | 4482 words | Yes (p = 0.04623) |
| 2. Readability | 42 Flesh Reading Ease | 41 Flesh Reading Ease | No (p= 0.2773) |
| 3. Complexity | 21% complex words | 22% complex words | Yes (p = 0.0003835) |
| 4. Lexical diversity | 0.25 TTR | 0.27 TTR | Yes (p = 0.004551) |
| 5. Number of references | 7235 words in References | 7419 words in References | No (p= 0.1799) |

## 4.3 Classification based on features only

Even though not all outcomes for the linguistic features were as expected, the results showed that the two types of papers did differ on several features. To see whether these features are indicative enough to distinguish retracted papers from non-retracted papers, based on solely the features, a classifier was used that used only the features as input. The performance on both the internal (Weighted F1-score of 0.60) and external (Weighted F1-score of 0.47) datasets were not good, as can be seen in figure 7. The recall, for the retracted papers, is for both internal and external datasets again higher than the precision. Also in this case, the classifier has a bias towards classifying papers as retracted. This bias is smaller than before, as the recall is not 1.00 for both datasets, but 0.66 and 0.64 for the retracted papers and can thus be considered to be a slight bias.

This performance is worse for the internal dataset than the scores of the Logistic Regression that was solely based on the pre-processed text. The performance on the external dataset is somewhat better than the performance of Logistic Regression, but still below chance. It means that features only are not enough to tell retracted papers apart from non-retracted papers. The performance of the internal dataset is again higher than the performance of the external dataset. The results are therefore, also in this case, not generalisable.

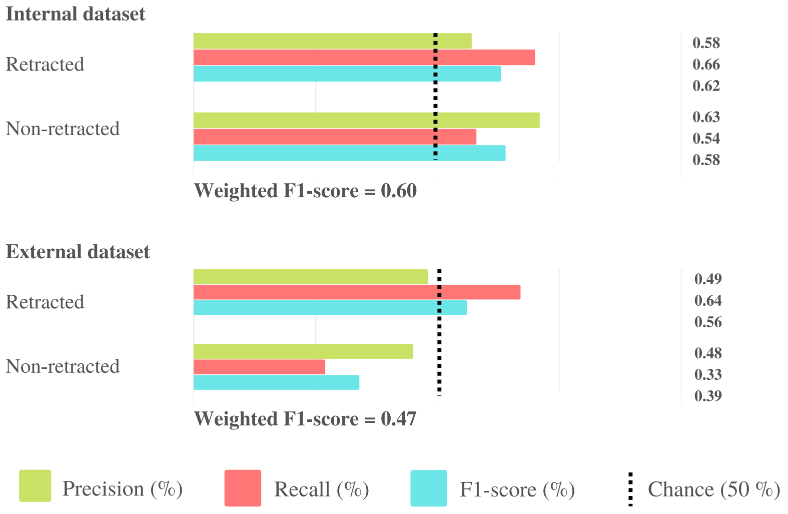


*Figure 7 Performance of Logistic Regression with linguistic features as input*

## 4.4 Classification based on text and features

Both classifiers are separately from each other not performing very well, especially not on the external dataset. As mentioned before, combining classifiers usually results in better performance. This is not the case for the stacked classifier used for this project. The performance of the stacked classifier is the exact same as the performance of the classifier that uses only text as input. The results can be found in the figure in Appendix 3. It could mean that the second classifier does not add any value to the stacked classifier. This could have been expected as the performance of the second classifier was similar to, and in fact even worse than, a random classifier, i.e. chance. The second classifier also had a lower Weighted F1-score on the internal dataset, than the first classifier had. Even though the performance on the external dataset was a bit higher for the second classifier, adding it together with the first classifier did not improve the overall performance on both the internal and external dataset.

The poor performance of the stacked classifiers provides an answer to the second research question [RQ1] 'Will including linguistic features result in better outcomes for classifying fraudulent research papers from non-fraudulent research papers?': adding features does not improve the performance of telling retracted papers apart from non-retracted papers.

An overview of the performance of all classifiers can be found in table 5.

*Table 5 Overview of the performance of the different classifiers*

| Classifier based on | Weighted F1-score internal dataset | Weighted F1-score external dataset |
|---|---|---|
| only text | 0.73 | 0.43 |
| only features | 0.60 | 0.47 |
| both text and features | 0.73 | 0.43 |

For all classifiers, the performance on the external dataset is lower than the performance on the internal dataset, when comparing the Weighted F1-scores. These findings are in line with the findings of the research by Franssen (2022) and (Lindenmeyer, 2022), who use similar data. This is not surprising as the journals that are used to train the model on, are different from the journals that are included in the external dataset. As mentioned before, the two journals in the external dataset are more general than the five journals in the internal dataset. This was done to check if the results are generalisable. It could however be that the model is fitted to specific topic words and is therefore performing worse on more general journals that include other topics.

An explanation for the lacking improvement of the performance, when including the features in the analysis, could be that the features are not different enough to tell deceptive language apart from truthful language. This was not expected as most averages and distributions of the data of these features were not completely alike. Statistical analyses, however, showed that two of the five features did not differ significantly, on whether they were retracted or not. It could be that the features were thus not too different from each other, for the classifier to be indicative.

Another explanation could come from the fact that in the retracted papers there is, besides fraudulent research, research with errors included. This kind of research does not try to cover up fraud by using deceptive language and will therefore will not differ on the linguistic features.

It could also be that the features were not operationalised in a correct way. This could be due to the fact that the results, from the analyses on the features, were not always in line with the literature. Another explanation for lacking improvement could be that the data still has too much noise and is therefore not representable. At last, it could also be the case that due to the high amount of research, that has been performed on deceptive language, researchers are aware of the potential markers in their language. It is possible that they know which features to pay attention to, that could prevent them from being caught on their misconduct. However, this not very likely as most of these features are results from processes and decisions that are made unaware by the authors.

## 4.3 Limitations

There are a few limitations of this project that should be addressed that could have influenced the results.

An important remark, that has to be made, is that it is not necessarily the case that every non-retracted paper is not fraudulent. As mentioned before, it is hard to put a number to how much research can be seen as fraudulent (see section 1.2.1.2). It is possible that in the non-fraudulent papers, there are also fraudulent papers of which its fraud has not (yet) been discovered. This would mean that the data is not completely reliable when it comes to the distinction of non-fraudulent language and fraudulent language. However, it is almost impossible to get such a big amount of papers on which a whole investigation is applied and of which it can be stated that the research is completely non-fraudulent.

Another remark, that should be mentioned, is that in this project I chose to use an equal distribution of retracted and non-retracted papers. In real life, the distribution of fraudulent and non-fraudulent research is unlikely to be exactly equal. One could argue that due to strict regulations, the expected amount of fraudulent research is lower than 50%. However, this cannot be stated with complete certainty as the exact numbers of fraudulent research are unknown. It could possibly be even higher, as research shows that researchers are hesitant to share their data for the sake of transparency (Brown & Heathers, 2017). Next to the fact that the majority of papers raises some type of problems, regarding their data, when investigating them.

As mentioned before, it is possible that there are no differences found in the linguistic features that point to deceptive language, due to the fact that the retracted papers are retracted for various reasons. Most reasons are due to intentional fraud, but there are also papers included that were retracted due to errors in the data. These kinds of papers are not trying to use deceptive language to mask the misconduct in the research. Therefore, the linguistic cues for deception could also not be present in the texts. Lindenmeyer (2022) shows that making the distinction between erroneous papers and fraudulent papers improves the performance of the classification models. It would be interesting to combine this research with the current project to see whether the linguistic features would improve the performance of the classifier, when the analysis is performed on solely fraudulent papers.

At last, it could be argued that the size of the dataset is still relatively small. With only 628 instances from seven journals, the dataset is not likely to be generalisable to academic papers from all existing journals. This was already seen by the fact that the models were not generalisable to the more general journals in the external dataset. It could be that the size of the dataset is too small to find general patterns of fraudulent research. It might also be interesting to see how the performance of the model would be if the models are trained on the more general papers.

# 5. Conclusion

The performance measures of the model, that uses only textual input, indicate that it is to some extent possible to distinguish retracted papers from non-retracted papers. The classifier is not generalisable, as the performance on the external dataset was even lower than chance. The classifier can thus only be used for papers within the same topic and/or journal.

As the goal of this project was to see whether including linguistic features, as markers for deceptive language, would improve the performance of the classifier, several features were analysed. These features were the quantity of lexicon, readability, complexity, lexical diversity and the number of references. Retracted papers include a significantly larger quantity of lexicons than non-retracted papers. While, on the other hand, non-retracted papers use significantly more complex words than retracted papers. The readability of the retracted papers is not significantly different than the non-retracted papers, as is the number of references. At last, the non-retracted papers have a higher Type Token Ratio, which means that they are more lexical diverse than retracted papers. Taken this all together, it can be stated that indeed some features show significant differences between the two types of papers.

If only these features are given as data to the classifier, the performance is somewhat comparable to the performance of the classifier on textual data. While it performed worse on the internal dataset, it performed better on the external dataset. The performance on the external dataset is however still below chance. Again, the results were not generalisable as the external dataset had also in this case a lower performance than the internal dataset. Due to the bad performance, the expectations were not high for combining the text and the linguistic features as input together for the classifier. Even though, prior research has shown that stacking classifiers results in better performance. These expectations for a lacking improvement were confirmed, as the performance of the model was exactly the same as the classifier which was solely based on the texts of the papers. It can thus be concluded, that adding the aforementioned linguistic features, does not improve the performance of the classifier for distinguishing retracted papers from non-retracted papers.

**Future work**

In I have chosen to use Logistic Regression as the classifier for distinguishing retracted from non-retracted papers. However, as mentioned before, there are several other classifiers. Some are just like Logistic Regression traditional methods for classification, such as Naïve Bayes. Others, such as BERT, are more modern. It would be interesting for future research to perform other kinds of classifiers on both text and features, to see whether the performance of these classifiers is similar to that of Logistic Regression.

As was shown before, the findings were not always in line with the literature, regarding the linguistic features that were ought to be markers for deceptive language. Future research is needed to see whether these linguistic features can be seen as markers, specifically for deceptive language in academic papers. Besides the five linguistic features that are used in this project, previous studies have shown that there are other linguistic markers that can also potentially detect fraud. Future research could focus on these other linguistic features and examine whether those are better predicters for distinguishing retracted papers from non-retracted ones.

# 6. References

Afroz, S., Brennan, M., & Greenstadt, R. (2012). Detecting hoaxes, frauds, and deception in writing style online. *2012 IEEE Symposium on Security and Privacy* (pp. 461-475). IEEE.

Alipourfard, N., Arendt, B., Benjamin, D. M., Benkler, N., Bishop, M., Burstein, M., . . . Wu, J. (n.d.). Systematizing confidence in open research and evidence (score). 2021.

Bachenko, J., Fitzpatrick, E., & Schonwetter, M. I. (2008). Verification and implementation of language-based deception indicators in civil and criminal narratives. *Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008)*, 41-48.

Brown, N. J., & Heathers, J. A. (2017). The GRIM test: A simple technique detects numerous anomalies in the reporting of results in psychology. *Social Psychological and Personality Science, 8(4)*, 363-369.

Burgoon, J. K., Blair, J. P., Qin, T., & Nunamaker, J. F. (2003). Detecting deception through linguistic analysis. *International Conference on Intelligence and Security Informatics*, 91-101.

Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. *Proceedings of the 23rd international conference on Machine learning*, (pp. 233-240).

Eleyan, D., Othman, A., & Eleyan, A. (2020). Enhancing software comments readability using flesch reading ease score. *Information, 11(9)*, 430.

Elgeldawi, E., Sayed, A., Galal, A. R., & Zaki, A. M. (2021). Hyperparameter tuning for machine learning algorithms used for arabic sentiment analysis. *Informatics (Vol. 8, No. 4)*, 79.

Franssen, J. (2022). Detecting Erroneous Research: Comparison of Shallow and Deep Learning Models.

Goel, S., & Gangolly, J. (2012). Beyond the numbers: Mining the annual reports for hidden cues indicative of financial statement fraud. *Intelligent Systems in Accounting, Finance and Management, 19(2)*, 75-89.

Gupte, A., Joshi, S., Gadgul, P., Kadam, A., & Gupte, A. (2014). Comparative study of classification algorithms used in sentiment analysis. *International Journal of Computer Science and Information Technologies, 5(5)*, 6261-6264.

Hancock, J. T., Curry, L. E., Goorha, S., & Woodworth, M. (2007). On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. *Discourse Processes, 45(1)*, 1-23.

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

Levitan, S. I., An, G., Wang, M., Mendels, G., Hirschberg, J., Levine, M., & Rosenberg, A. (2015). Cross-cultural production and detection of deception from speech. *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection*, 1-8.

Lindenmeyer, A. (2022). Classification of retracted and non-retracted scientific articles.

Markowitz, D. M., & Hancock, J. T. (2016). Linguistic obfuscation in fraudulent science. *Journal of Language and Social Psychology, 35(4)*, 435-445.

Mbaziira, A., & Jones, J. (2016). A text-based deception detection model for cybercrime. *Int. Conf. Technol. Manag.*

National Academy of Sciences, National Academy of Engineering (US) and Institute of Medicine (US) Committee on Science, Engineering, and Public Policy. (2009). On Being a Scientist: A Guide to Responsible Conduct in Research: Third Edition. Washington D.C., United States of America: National Academies Press (US).

Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003). Lying words: Predicting deception from linguistic styles. *Personality and social psychology bulletin, 29(5)*, 665-675.

Porter, S., & Yuille, J. C. (1996). The language of deceit: An investigation of the verbal clues to deception in the interrogation context. *Law and Human Behavior, 20(4)*, 443-458.

Powers, D. M. (2020). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*.

Pupovac, V., & Fanelli, D. (2015). Scientists admitting to plagiarism: a meta-analysis of surveys. . *Science and engineering ethics, 21(5)*, 1331-1352.

Sarewitz, D. (2016). The pressure to publish pushes down quality. *Nature, 533(7602)*, 147.

Schumm, W. R., Crawford, D. W., & Lockett, L. (2019). Using statistics from binary variables to detect data anomalies, even possibly fraudulent research. *Psychology Research and Applications, 1(4)*, 112-118.

Simmons, J., Nelson, L. D., & Simonsohn, U. (2013). Life After P-Hacking. *Meeting of the society for personality and social psychology, New Orleans, LA*, 17-19.

Smith, R. (2005). Investigating the previous studies of a fraudulent author. *BMJ, 331(7511)*, 288-291.

Stretton, S., Bramich, N. J., Keys, J. R., Monk, J. A., Ely, J. A., Haley, C., . . . Woolley, K. L. (2012). Publication misconduct and plagiarism retractions: a systematic, retrospective study. *Current medical research and opinion, 28(10)*, 157501583.

Zhou, L., & Sung, Y. W. (2008). Cues to deception in online Chinese groups. *Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008)* (p. 146). IEEE.

# 7. Appendix

## Appendix 1: Notebook scripts

The Python Notebooks that are used for this project, can be retrieved via my GitHub repository: https://github.com/schmidteveline/thesis

Here the following Python Notebooks can be found:

- Seven notebooks for all the pre-processing steps
  - STEP1_scraper_non_retracted.ipynb
  - STEP1_scraper_retracted.ipynb
  - STEP1_pdf_to_text_non_retracted.ipynb
  - STEP1_pdf_to_text_retracted.ipynb
  - STEP2_Creating Balanced Dataset.ipynb
  - STEP3_preprocessing part 1.ipynb
  - STEP4_preprocessing part 2.ipynb
- One notebook for the analysis of the five linguistic features
  - linguistic_feature_analyses.ipynb
- One notebook for the Logistic Regression classifiers
  - logistic_regression_text_and_features.ipynb

# Appendix 2: Statistical analyses on linguistic features

Overview of the results of the statistical analyses for all five linguistic features.

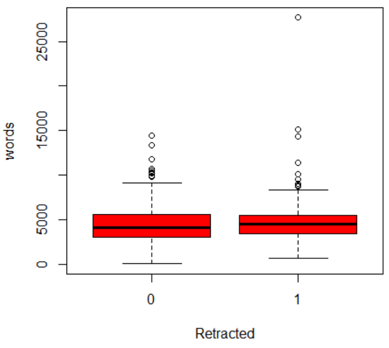
## Analysis on linguistic features - Thesis ADS

## Download data

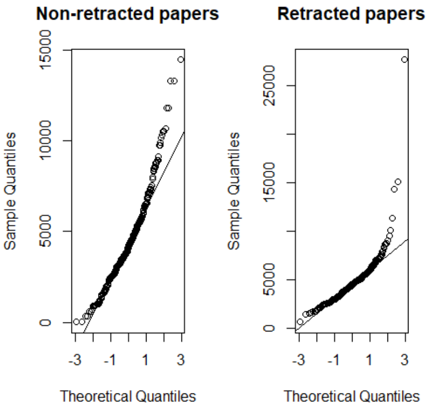
```
dat7 <- read.csv(file = 'df_numerical_seven_journals_features_vf.csv')
table(dat7$Retracted)

##
##   0   1
## 314 314
```

## Quantity of lexicon

```
boxplot(words~Retracted, dat=dat7, col = 'red')
```



```
par(mfrow=c(1,2))
qqnorm(dat7[dat7$Retracted==0,]$words, main='Non-retracted papers')
qqline(dat7[dat7$Retracted==0,]$words)
qqnorm(dat7[dat7$Retracted==1,]$words, main='Retracted papers')
qqline(dat7[dat7$Retracted==1,]$words)
```



```
shapiro.test(dat7[dat7$Retracted==0,]$words) #significant, so not normally distributed

##
##  Shapiro-Wilk normality test
##
## data:  dat7[dat7$Retracted == 0, ]$words
## W = 0.94214, p-value = 9.613e-10

shapiro.test(dat7[dat7$Retracted==1,]$words) #significant, so not normally distributed
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dat7[dat7$Retracted == 1, ]$words
## W = 0.72595, p-value < 2.2e-16
```

```
wilcox.test(words ~ Retracted, dat=dat7, alternative='two.sided') # p < 0.05, significant d
ifferent
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  words by Retracted
## W = 44766, p-value = 0.04623
## alternative hypothesis: true location shift is not equal to 0
```

```
library(effsize)
```

```
## Warning: package 'effsize' was built under R version 4.0.5
```

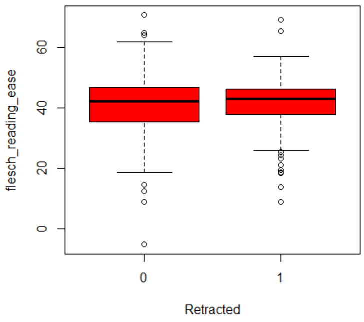
```
cliff.delta(words ~ Retracted, dat=dat7)
```

```
##
## Cliff's Delta
##
## delta estimate: -0.09193071 (negligible)
## 95 percent confidence interval:
##         lower         upper
## -0.1815016491 -0.0008469544
```

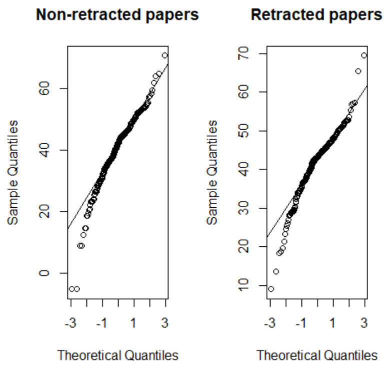
I assessed if retracted papers have a different quantity of lexicon than non-retracted papers. My $H_0: P(R > NR) = P(NR > R)$, and our $H_a: P(R > NR) \neq P(NR > R)$. There are 314 retracted and 314 non-retracted papers. The visualization is shown above and shows that the retracted papers appeared to have a very small higher quantity of lexicon. The quantile-quantile plot shows that the distribution of both groups did not seem normal (this was validated via a Shapiro-Wilk test for normality). Due to this, we used a Mann-Whitney U test (one-tailed) showing that the difference was significant at the 0.05 $\alpha$-level. The $p$-value of the Mann-Whitney U test (with U = 44766 was 0.04623. The effect size Cliff's $d$ = -0.09193071. We accept the alternative hypothesis that retracted papers have a different quantity of lexicon than non-retracted papers.

## Flesch Reading Ease

```
boxplot(flesch_reading_ease~Retracted, dat=dat7, col = 'red')
```



```
par(mfrow=c(1,2))
qqnorm(dat7[dat7$Retracted==0,]$flesch_reading_ease, main='Non-retracted papers')
qqline(dat7[dat7$Retracted==0,]$flesch_reading_ease)
qqnorm(dat7[dat7$Retracted==1,]$flesch_reading_ease, main='Retracted papers')
qqline(dat7[dat7$Retracted==1,]$flesch_reading_ease)
```

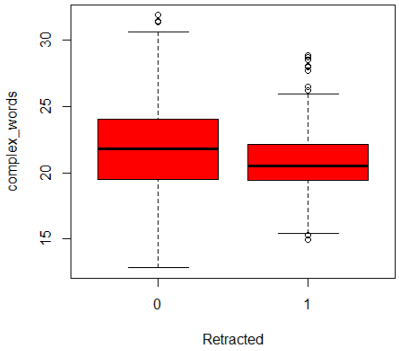


```
shapiro.test(dat7[dat7$Retracted==0,]$flesch_reading_ease) #significant, so not normally di
stributed

##
##  Shapiro-Wilk normality test
##
## data:  dat7[dat7$Retracted == 0, ]$flesch_reading_ease
## W = 0.95604, p-value = 4.21e-08

shapiro.test(dat7[dat7$Retracted==1,]$flesch_reading_ease) #significant, so not normally di
stributed

##
##  Shapiro-Wilk normality test
##
## data:  dat7[dat7$Retracted == 1, ]$flesch_reading_ease
## W = 0.94495, p-value = 1.967e-09

wilcox.test(flesch_reading_ease ~ Retracted, dat=dat7, alternative='two.sided') # p < 0.05,
significant different

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  flesch_reading_ease by Retracted
## W = 46828, p-value = 0.2773
## alternative hypothesis: true location shift is not equal to 0

library(effsize)
cliff.delta(flesch_reading_ease ~ Retracted, dat=dat7)

##
## Cliff's Delta
##
## delta estimate: -0.05010345 (negligible)
## 95 percent confidence interval:
##       lower       upper
## -0.14089914  0.04152803
```

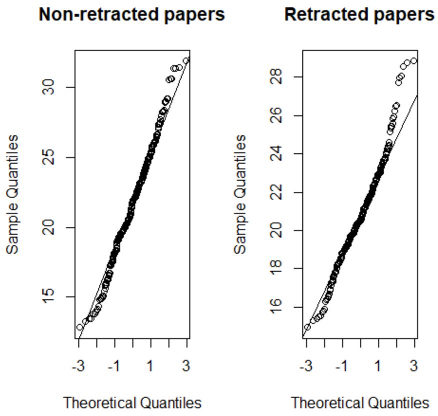
I assessed if retracted papers have a different quantity of lexicon than non-retracted papers. My $H_0: P(R > NR) = P(NR > R)$, and our $H_a: P(R > NR) \neq P(NR > R)$. There are 314 retracted and 314 non-retracted papers. The visualization is shown above and shows that the retracted papers appeared to have a similar distribution of flesch reading ease. The quantile-quantile plot shows that the distribution of both groups did not seem normal (this was validated via a Shapiro-Wilk test for normality). Due to this, we used a Mann-Whitney U test (one-tailed) showing that the difference was not significant at the 0.05 $\alpha$-level. The $p$-value of the Mann-Whitney U test (with U = 46828 was 0.2773 The effect size Cliff's $d$ = -0.05010345. We accept the null hypothesis that retracted papers have no different flesch reading ease than non-retracted papers.

## Complexity

```
boxplot(complex_words~Retracted, dat=dat7, col = 'red')
```



```
par(mfrow=c(1,2))
qqnorm(dat7[dat7$Retracted==0,]$complex_words, main='Non-retracted papers')
qqline(dat7[dat7$Retracted==0,]$complex_words)
qqnorm(dat7[dat7$Retracted==1,]$complex_words, main='Retracted papers')
qqline(dat7[dat7$Retracted==1,]$complex_words)
```



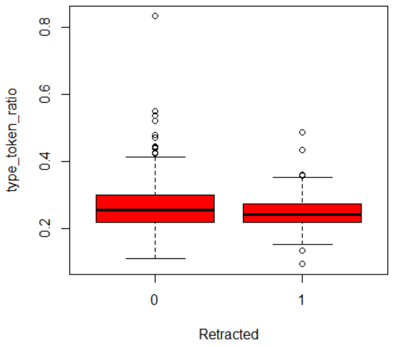
```
shapiro.test(dat7[dat7$Retracted==0,]$complex_words) #in-significant, so normally distribut
ed

##
##  Shapiro-Wilk normality test
##
## data:  dat7[dat7$Retracted == 0, ]$complex_words
## W = 0.99266, p-value = 0.1257

shapiro.test(dat7[dat7$Retracted==1,]$complex_words) #significant, so not normally distribu
ted

##
##  Shapiro-Wilk normality test
##
## data:  dat7[dat7$Retracted == 1, ]$complex_words
## W = 0.97572, p-value = 3.732e-05

wilcox.test(complex_words ~ Retracted, dat=dat7, alternative='two.sided') # p < 0.05, signi
ficant different

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  complex_words by Retracted
## W = 57372, p-value = 0.0003835
## alternative hypothesis: true location shift is not equal to 0
```

```
library(effsize)
cliff.delta(complex_words ~ Retracted, dat=dat7)

##
## Cliff's Delta
##
## delta estimate: 0.1637693 (small)
## 95 percent confidence interval:
##      lower      upper
## 0.07171047 0.25306155
```

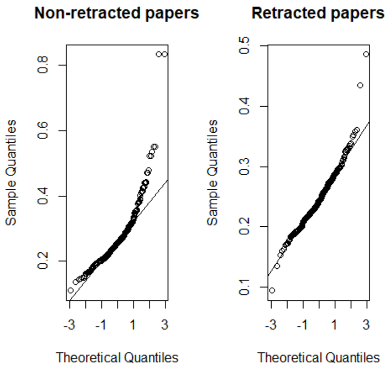
I assessed if retracted papers have a different quantity of lexicon than non-retracted papers. My $H_0: P(R > NR) = P(NR > R)$, and our $H_a: P(R > NR) \neq P(NR > R)$. There are 314 retracted and 314 non-retracted papers. The visualization is shown above and shows that the retracted papers appeared to have lower amount of complex words than non-retracted papers. The quantile-quantile plot shows that the distribution of one group did not seem normal (this was validated via a Shapiro-Wilk test for normality). Due to this, we used a Mann-Whitney U test (one-tailed) showing that the difference was significant at the 0.05 $\alpha$-level. The $p$-value of the Mann-Whitney U test (with U = 57372 was 0.0003835. The effect size Cliff's $d$ = 0.1637693. We accept the alternative hypothesis that retracted papers have a different amount of complex words than non-retracted papers.

## Lexical Diversity

```
boxplot(type_token_ratio~Retracted, dat=dat7, col = 'red')
```



```
par(mfrow=c(1,2))
qqnorm(dat7[dat7$Retracted==0,]$type_token_ratio, main='Non-retracted papers')
qqline(dat7[dat7$Retracted==0,]$type_token_ratio)
qqnorm(dat7[dat7$Retracted==1,]$type_token_ratio, main='Retracted papers')
qqline(dat7[dat7$Retracted==1,]$type_token_ratio)
```



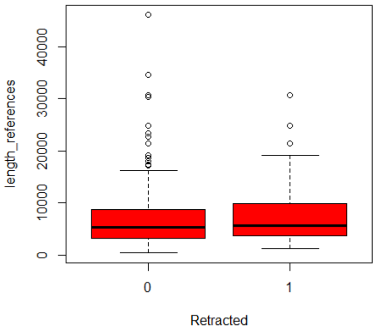
```
shapiro.test(dat7[dat7$Retracted==0,]$type_token_ratio) #significant, so not normally distr
ibuted

##
##  Shapiro-Wilk normality test
##
```

```
## data:  dat7[dat7$Retracted == 0, ]$type_token_ratio
## W = 0.82273, p-value < 2.2e-16

shapiro.test(dat7[dat7$Retracted==1,]$type_token_ratio) #significant, so not normally distr
ibuted

##
##  Shapiro-Wilk normality test
##
## data:  dat7[dat7$Retracted == 1, ]$type_token_ratio
## W = 0.96059, p-value = 1.689e-07

wilcox.test(type_token_ratio ~ Retracted, dat=dat7, alternative='two.sided') # p < 0.05, si
gnificant different

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  type_token_ratio by Retracted
## W = 55749, p-value = 0.004551
## alternative hypothesis: true location shift is not equal to 0

library(effsize)
cliff.delta(type_token_ratio ~ Retracted, dat=dat7)

##
## Cliff's Delta
##
## delta estimate: 0.1308471 (negligible)
## 95 percent confidence interval:
##      lower      upper
## 0.03968646 0.21984753
```

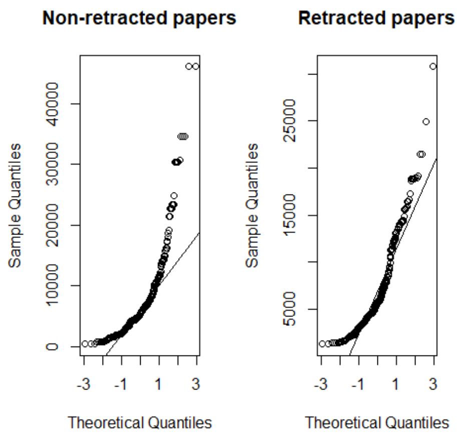
I assessed if retracted papers have a different quantity of lexicon than non-retracted papers. My $H_0\colon P(R > NR) = P(NR > R)$, and our $H_a\colon P(R > NR) \neq P(NR > R)$. There are 314 retracted and 314 non-retracted papers. The visualization is shown above and shows that the retracted papers appeared to have lower lexical diversity than non-retracted papers. The quantile-quantile plot shows that the distribution of both groups did not seem normal (this was validated via a Shapiro-Wilk test for normality). Due to this, we used a Mann-Whitney U test (one-tailed) showing that the difference was significant at the 0.05 $\alpha$-level. The $p$-value of the Mann-Whitney U test (with U = 55749 was 0.004551. The effect size Cliff's $d$ = 0.1308471. We accept the alternative hypothesis that retracted papers have a different lexical diversity than non-retracted papers.

## Number of references

```
boxplot(length_references~Retracted, dat=dat7, col = 'red')
```



```
par(mfrow=c(1,2))
qqnorm(dat7[dat7$Retracted==0,]$length_references, main='Non-retracted papers')
qqline(dat7[dat7$Retracted==0,]$length_references)
qqnorm(dat7[dat7$Retracted==1,]$length_references, main='Retracted papers')
qqline(dat7[dat7$Retracted==1,]$length_references)
```



```
shapiro.test(dat7[dat7$Retracted==0,]$length_references) #significant, so not normally distributed

##
##  Shapiro-Wilk normality test
##
## data:  dat7[dat7$Retracted == 0, ]$length_references
## W = 0.73828, p-value < 2.2e-16

shapiro.test(dat7[dat7$Retracted==1,]$length_references) #significant, so not normally distributed

##
##  Shapiro-Wilk normality test
##
## data:  dat7[dat7$Retracted == 1, ]$length_references
## W = 0.88007, p-value = 5.868e-15

wilcox.test(length_references ~ Retracted, dat=dat7, alternative='two.sided') # p > 0.05, not significant different

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  length_references by Retracted
## W = 46249, p-value = 0.1799
## alternative hypothesis: true location shift is not equal to 0

library(effsize)
cliff.delta(length_references ~ Retracted, dat=dat7)

##
## Cliff's Delta
##
## delta estimate: -0.06185849 (negligible)
## 95 percent confidence interval:
##       lower       upper
## -0.15150626  0.02879861
```

I assessed if retracted papers have a different quantity of lexicon than non-retracted papers. My $H_0: P(R > NR) = P(NR > R)$, and our $H_a: P(R > NR) \neq P(NR > R)$. There are 314 retracted and 314 non-retracted papers. The visualization is shown above and shows that the retracted papers appeared to have higher amount of references than non-retracted papers. The quantile-quantile plot shows that the distribution of both groups did not seem normal (this was validated via a Shapiro-Wilk test for normality). Due to this, we used a Mann-Whitney U test (one-tailed) showing that the difference was not significant at the 0.05 $\alpha$-level. The $p$-value of the Mann-Whitney U test (with U = 46249 was 0.1799. The effect size Cliff's $d$ = -0.06185849. We accept the null hypothesis that retracted papers do not have a different number of references than non-retracted papers.

## Appendix 3: Performance classifier based on text and linguistic features

Figure 8 shows the performance of the Logistic Regression that has text and linguistic features as input. The performance is exactly the same as the performance that uses only textual input.

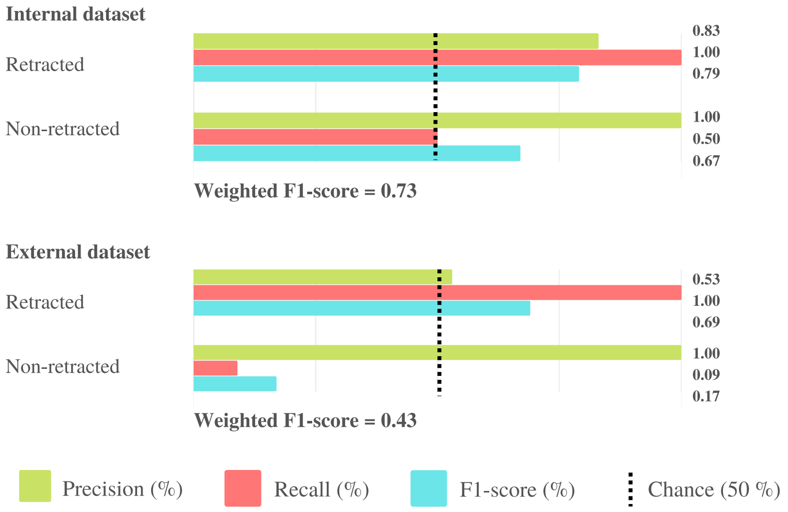


*Figure 8 Performance of Logistic Regression with both texts and linguistic features as input*