Detecting Scientific Fraud Using Argument Mining

Gabriel Freedman, Francesca Toni
Imperial College London, UK
{g.freedman22,f.toni}@imperial.ac.uk

Abstract

A proliferation of fraudulent scientific research in recent years has precipitated a greater interest in more effective methods of detection. There are many varieties of academic fraud, but a particularly challenging type to detect is the use of paper mills and the faking of peer-review. To the best of our knowledge, there have so far been no attempts to automate this process. The complexity of this issue precludes the use of heuristic methods, like pattern-matching techniques, which are employed for other types of fraud. Our proposed method in this paper uses techniques from the Computational Argumentation literature (i.e. argument mining and argument quality evaluation). Our central hypothesis stems from the assumption that articles that have not been subject to the proper level of scrutiny will contain poorly formed and reasoned arguments, relative to legitimately published papers. We use a variety of corpora to test this approach, including a collection of abstracts taken from retracted papers. We show significant improvement compared to a number of baselines, suggesting that this approach merits further investigation.

1 Introduction

The growing problem of fraudulent academic research poses a threat to scientific progress. Research is an iterative process, where arguments presented in previous papers are used as the basis of subsequent work. Researchers do not necessarily have the time or resources available to verify that all the claims that have been made in previous publications are well-formulated, or valid. Instead they tend to rely on the scrutiny imposed during the publication process to offer reasonable guarantees about the legitimacy of the content.

However, these guarantees have been undermined by revelations about the extent of malpractice taking place in many scientific publications (Cookson, 2023; Economist, 2023). As an indication of the scale of the problem, there have been estimates that over a fifth of new medical publications are implicated in some form of fraudulent activity (Sabel et al., 2023). Prevalent types of fraud include: plagiarism, data manipulation and circumvention of a legitimate peer-review process.

Relatively simple methods can help detect some of these forms of fraud, such as pattern matching in the case of plagiarism (Butler, 2018). However, where the peer-review process is compromised, or the services of paper mills are employed, these techniques are not effective. In these cases, researchers have had some success detecting characteristic patterns (van Noorden, 2021; Else and van Noorden, 2021) - but this is not a universal panacea as it is simple for authors to make slight stylistic alterations to avoid these methods of detection. The consensus of the research community investigating these matters is that there are a large number of such articles that have not yet been retracted from the journals they are published in, and this number is on the rise (Sabel and Seifert, 2021).

The hypothesis underlying this research is that papers published by illicit means (specifically those that have been produced by paper mills, or have not undergone peer-review) will be based on substandard reasoning. This could take the form of fallacious arguments. Alternatively, arguments may be based on spurious premises, or lack any relevant and novel insights. We believe this is a legitimate supposition due to the nature of scientific inquiry: the fundamental aim of scientific research is to construct valid and interesting arguments from a sound empirical or theoretical basis.

Anecdotal evidence supporting this hypothesis is presented in Table 1. Both arguments address the efficacy of social distancing as a method to combat the spread of COVID-19. However, the argument given in the retracted article is very weak. The claim: 'Social distancing measures ought to be followed by everyone to minimize the spread of COVID-19', is perfectly reasonable. The premises

Proceedings of the 11th Workshop on Argument Mining, pages 15–28

Retracted	Not retracted
Social distancing measures ought to be followed by everyone to minimize the spread of COVID-19. Eventually, maintaining social distance will become a habit in the future. Owing to that, our proposed system gives an accurate output of 90% at detecting people with a one-meter distance between them in public areas, which also provides indications in green and red bounding boxes around people.	After three COVID-19 waves, the growing number of new infections still reminds us of the importance of taking precautionary measures. SD and wearing masks have been proven to be efficient nonpharmaceutical intervention measures (Özbek, Syed, & Öksüz, 2021). They are low-cost, convenient, and noninvasive to slow the spread of COVID-19 and flatten the curves of infection (Srivastava, Zhao, Manay, & Chen, 2021).

Table 1: Comparable arguments for social distancing presented in a retracted article (Pooranam et al., 2021), and non-retracted article (Himeur et al., 2022).

that follow, however, are very loosely connected to the claim, and the argument in its entirety is both unconvincing and hard to follow. On the other hand, the argument in the non-retracted article is much more effective. The premises are directly addressing the points raised in the claim, and they back up their assertions with verifiable evidence in the form of citations.

In order to empirically test this hypothesis, we adopt techniques from the Computational Argumentation literature. Specifically, we build on past research in the fields of **argument mining** (Lawrence and Reed, 2020) and **argument quality evaluation** (Toledo et al., 2019). Models trained for these purposes are able to extract arguments from a passage of text and evaluate the quality of such arguments respectively. When done sequentially, this amounts to a way of assessing the reasoning present in a piece of text.

We use a number of pre-existing corpora both to train and test the various models we develop. These include datasets that have been compiled especially for scientific argument mining: SciARK (Fergadis et al., 2021) for training and AbstRCT (Mayer et al., 2020) for testing. Also, we use a dataset that contains human evaluated arguments for training our argument quality evaluation model: the Grammarly Argument Quality Corpus (GAQCorpus) (Lauscher et al., 2020).

In order to collect a sufficient sample of retracted articles to evaluate the performance of the complete system, we use the Retraction Watch database (Marcus and Oransky, 2023). The metadata included therein allows us to specify the subset of retracted articles that we are interested in detecting.

Our initial results indicate that implementing this strategy leads to a considerable improvement in detecting fraudulent articles, compared with a number of baselines. This suggests that the developed method has theoretical validity and merits further investigation.

2 Related Work

2.1 Scientific Fraud Detection

As awareness grows about the existing and potential problems caused by academic fraud (Bolland et al., 2022; Fanelli et al., 2022; Kim et al., 2019; Garmendia et al., 2019), researchers have begun to take steps to tackle the problem. There is generally still an emphasis put on human-centered interventions. Such proposals include introducing more stringent criteria for publications to choose their referees (Mavrogenis and Scarlat, 2023), improving the quality of oversight and guidance offered by regulatory bodies (Candal-Pedreira et al., 2021), and producing effective guidelines to help both academics and journals cooperate to avoid any fraudulent activity (Wager et al., 2017).

Due to the scale of the problem, some researchers have recognised that it is necessary to at least partially automate the discovery process. There have been varying degrees of automation suggested. Zhao et al. (2021) propose a method to improve the selection of referees. They compare a vector embedding of the paper under review with embedded representations of a number of potential referees' previous papers to more accurately determine who has the most relevant expertise.

Other approaches focus on using information about the authors of the papers or the publication venue itself. Abalkina has proposed using the archives (Abalkina, 2021a) and the metadata (Abalkina, 2021b) of papers that have appeared in compromised journals in order to detect other publication venues that may have also been compromised. Similarly Chakraborty et al. (2021) focus on analysing irregular citation patterns to find self-referencing networks of fraudulent papers.

Some authors have attempted to propose solutions that rely more fully on computational meth-

ods. Haunschild and Bornmann (2021) investigate the possibility of using scepticism expressed on social media as a metric to determine possible fraudulent activity. Furthermore, Kinney et al. (2021) use measures of text overlap to detect plagiarism and Horton et al. (2020) attempt to use statistical methods to uncover patterns in manipulated data. However, these latter two approaches are not applicable to all types of academic fraud, and would not necessarily work for faked peer-reviews or papers produced by paper mills.

We aim to advance the current state of this research by developing and implementing a system that can make fully automated predictions about whether a paper has bypassed a legitimate peer-review, or equivalently has been produced by a paper mill. We were not able to find any comparable research in the literature, suggesting that our work constitutes a novel research program.

2.2 Argument Mining

Argument mining (Lawrence and Reed, 2020) is an important task in Computational Argumentation. It is the automatic extraction of arguments contained within text. Once these arguments have been identified, it is possible to create formal representations which deliver a greater flexibility and ability to reason (Peldszus and Stede, 2013).

The task is a very challenging aspect of natural language processing, and has not yet been solved with a high level of accuracy. The heterogeneity of argument types and structure make reliable and consistent representations hard to achieve. However, since the advent of the transformer architecture (Vaswani et al., 2017) and the consequent improvement in language modelling capabilities (Devlin et al., 2019; Brown et al., 2020), advances have also been made in the field of argument mining. The ability to fine-tune pre-trained large language models (LLMs) on task-specific datasets has made the integration of argument mining into practical applications a possibility.

There are roughly three subtasks that make up the argument mining task: the detection of individual argumentative entities (e.g. premises and claims), intra-argument relations (how premises and claims in arguments relate) and inter-argument relations (how different arguments relate).

There are a number of specifically curated datasets for each of these tasks. Stab and Gurevych (Stab and Gurevych, 2016) demonstrate the feasibility of developing guidelines that lead to a high inter-annotator agreement, producing a corpus of over four hundred annotated persuasive essays. There are also a number of datasets with a particular focus on the scientific domain. These include SciARG (Accuosto et al., 2021), SCiARK (Fergadis et al., 2021) and AbstRCT (Mayer et al., 2020).

Modelling the distribution of entities (claims and premises), and modelling the relationships between these entities are often separated into distinct tasks. For example, Cocarascu et al. (2020) develop a set of domain-agnostic models that can be applied to the relation prediction task. Ruiz-Dolz et al. (2020) attempt to solve a similar task, focusing on comparing the performance of different transformer-based architectures.

Similarly to our work, Fergadis et al. (2021) develop a variety of models that specifically address the entity identification subtask. Furthermore, Mayer et al. (2020) and Accuosto et al. (2021) both develop two types of model, one for the entity identification task and one for the relation prediction task. Thorburn and Kruger (2022), on the other hand, test different optimisation techniques with a GPT-like model, to attempt to create a more adaptable and versatile approach to different argument mining subtasks.

2.3 Argument Quality Evaluation

Evaluating the quality of arguments is a relatively unexplored aspect of Computational Argumentation. Wachsmuth et al. (2017) set out a broad framework that can be used to help define argument quality.

There have been a number of practical efforts to compile such a dataset. Initial efforts used a pairwise comparison between arguments (Habernal and Gurevych, 2016; Simpson and Gurevych, 2018). This is the most straightforward approach for annotators but is limited in its applicability to multiple arguments in different domains.

As a part of the IBM *Project Debater*, this approach was refined (Toledo et al., 2019; Gretz et al., 2019). In order to produce arguments with continuous numerical quality representations, questions with binary answers were asked about each argument. Numerous annotators were asked to consider the same arguments. Various methods of taking a weighted average are then explored, providing a continuous quality for each argument between 0 and 1. Joshi et al. (2023) have recently compiled a similar dataset, but they include 'argument-analysis pairs', which provide additional rationale behind

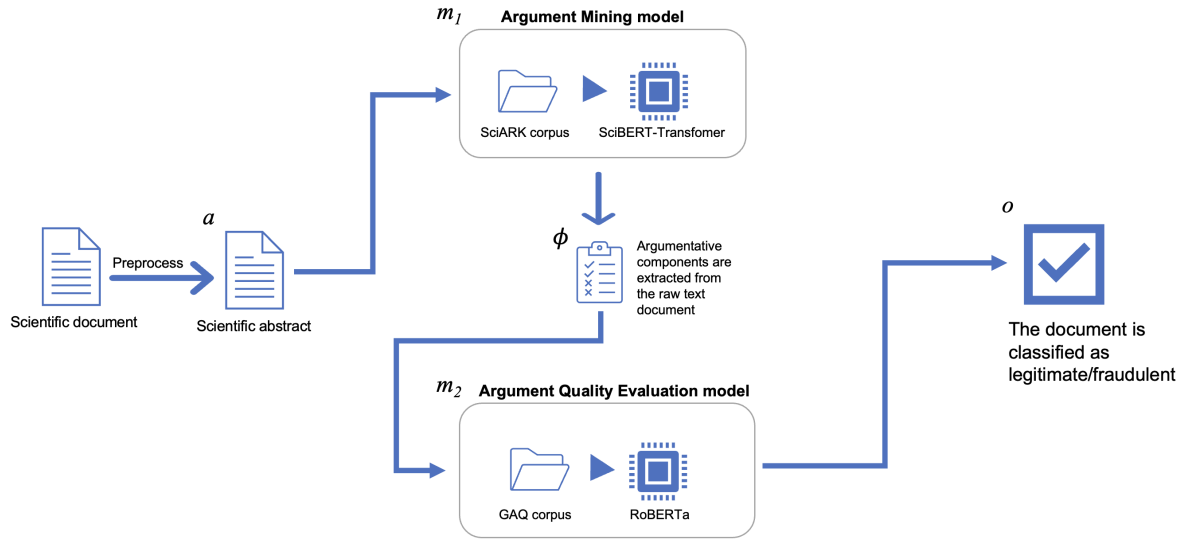

Figure 1: An overview of our proposed system architecture.

why the argument is effective.

Lauscher et al. (2020) take a more direct approach to producing continuous quality ratings. They take the average of three annotators' ratings, on a scale of one to five, over three different measures of quality. This allows for a more descriptive and fine-grained interpretation of different aspects of argument quality. We leverage this innovation by using state-of-the-art natural language processing techniques to develop an effective model for argument quality prediction.

3 Methodology

The goal of this research is to test the hypothesis that evaluating the quality of a scientific article's reasoning can be an effective way of determining whether it was produced fraudulently. In practice, this amounts to extracting and then evaluating arguments that are indicative of the overall level of reasoning present in the article. We achieve this by training two models separately.

An important feature of our framework is that we only analyse the abstracts of papers. We have two primary motivations for this decision. Firstly, it provides an effective way to minimise artefacts in our data. Research papers are generally heterogeneous in format, containing inconsistently structured sections. Abstracts, however, are fairly consistent in form, and from an argumentative perspective, usually contain the principle argument motivating the rest of the paper.

The second reason is that if we are able to demonstrate that the abstract alone is sufficient, then this offers practical advantages. Only considering a minimal subset of the entire text reduces both the theoretical and computational burden. This latter consideration is especially pertinent when considering that deploying such a system would be most advantageous in situations which require processing high volumes of inputs.

We describe our proposed fraud prediction framework with the following formalisation.

Let $a = \{s_1, s_2, ..., s_n\}$ represent a paper abstract consisting of a list of n sentences. We define a model m_1 such that:

$$m_1 : s_i \mapsto c_j \tag{1}$$

where $c_j \in \{none, evidence, claim\}$ is the category to which the sentence s_i is mapped.

The list of categorisations $c = \{c_1, c_2, ..., c_n\}$ along with the original abstract a is then transformed using a transformation function ϕ:

$$\phi : (a, \{c_1, c_2, ..., c_n\}) \mapsto t \tag{2}$$

where t is the transformed representation of the abstract a, amounting to a single string.

The transformed abstract is then input to another model m_2:

$$m_2 : t \mapsto v \tag{3}$$

where $v \in [1, 5]$ is a real-valued output.

The final status o of the abstract is determined by comparing the output v to a threshold $T = 3$:

$$o = \begin{cases} \text{legitimate}, & \text{if } v > T \\ \text{fraudulent}, & \text{if } v \leq T \end{cases} \quad (4)$$

Both models in our system architecture (see Figure 1), are realised based on fine-tuned foundation models (Bommasani et al., 2021), using data from a number of different sources. In Section 3.1 we describe the data we use for fine-tuning and testing our individual models, as well as testing our system in its entirety. Sections 3.2 and 3.3 provide descriptions of the Argument Mining (AM) model (m_1) and Argument Quality Evaluation (AQE) model (m_2) respectively. Finally in Section 3.4, we outline the entire system as presented in Figure 1.

3.1 Data

There is a relatively limited amount of annotated data focused on the tasks comprising argument mining, especially those that are relevant to scientific domains. Likewise, there have been even fewer datasets compiled for the purpose of training argument quality evaluation models. However, there is a sufficient amount of data available to support the training and testing of the models required to realise our system.

The SciARK (Fergadis et al., 2021) dataset consists of 1,000 annotated scientific abstracts (containing 12,374 annotated sentences in total), across a range of different domains (each domain corresponding to a UN Sustainable Development Goal (Biermann et al., 2017)). For each abstract, every sentence has been annotated with one of three labels: *evidence* (equivalent to premise), *claim* and *neither*. This dataset is used to train the AM model.

We use the AbstRCT dataset (Mayer et al., 2020) for evaluating the performance both of our AM model, and our fraud prediction system in its entirety. The dataset consists of 669 abstracts, and is annotated in the same way as the SciARK dataset, with each sentence being labelled either *evidence*, *claim* or *neither*. The abstracts are taken from publications in prestigious peer-reviewed medical journals. The rigorous and scientific form of argumentation contained in these samples makes them well-suited for testing the capabilities of the AM model we developed. Furthermore, the quality of the journals chosen by the compilers provides us with a sufficient level of confidence that none of the samples contained within the dataset were produced using fraudulent means. Therefore, it is also suitable to be used as the representative sample of *legitimate abstracts* that we use to evaluate our complete fraud prediction system.

For fine-tuning our AQE model we take 1,104 samples from GAQCorpus, compiled by (Lauscher et al., 2020). This dataset makes use of arguments taken from various internet forums. Annotators were recruited to give each argument a score on a scale between one and five, for each of three metrics: *cogency*, *effectiveness* and *reasonableness*. Despite the subjectivity inherent in human judgements, taking the average of multiple annotators' scores for each sample reduces the amount of noise present in the data.

In order to compile a sufficient corpus of *fraudulent abstracts*, for evaluating the performance of our system, we collected 420 relevant papers from the Retraction Watch database (Marcus and Oransky, 2023). The database currently contains tens of thousands of retracted articles from a wide variety of different journals. However, there are numerous reasons cited for each retraction, with the majority having to do with fake data or plagiarism.

For the sake of this study we are only interested in the subset of papers that have been published either by means of a paper mill, or by faking the peer-review. Furthermore, we restrict the papers we collect to the medical domain, in order to match the domain of those contained in the AbstRCT dataset, which constitute our test set of *legitimate abstracts*. We do this so our results are not influenced by features that are irrelevant to the focus of our study - namely the subject of the papers.

3.2 Argument Mining model

The AM model we developed is influenced by the architecture in (Fergadis et al., 2021). The model consists of three components: a *Sentence Encoder*, *Context Encoder* and a *Fully Connected Layer*.

The Sentence Encoder is a SciBERT model (Beltagy et al., 2019) - a BERT-like LLM, which has been trained specifically to improve performance on scientific texts. For each sentence in the input text, a `[CLS]` token is outputted, representing a sentence vector $s \in \mathbb{R}^{728}$. These tokens are used as input to the Context Encoder, providing a representation of the entire abstract during the production of the embedding for each sentence.

The Context Encoder provides a detailed representation of the specific sentence being consid-

ered. The best performing implementation from (Fergadis et al., 2021) uses a BiLSTM (Graves and Schmidhuber, 2005), taking as input both the sentence vectors before and after the current sentence. The dense layer simply takes the embedded representation and returns an output of one of three categories: *evidence*, *claim* or *null*.

We augment this model by replacing the BiLSTM Context Encoder with a transformer, better suited for handling long-range dependencies, which is particularly useful in the context of a scientific abstract where all concepts mentioned are often relevant throughout the entire passage.

Due to the nature of the setting we forego inter-argument and intra-argument relation prediction. This is because all our samples are scientific abstracts. These are relatively short passages, and also, in theory, should only be presenting the one principle argument being introduced in the paper. This means that simply identifying the argumentative entities is sufficient, as we assume that each sample contains at most one claim (possibly spanning multiple sentences), and potentially multiple premises supporting that claim.

3.3 Argument Quality Evaluation model

Once the arguments have been extracted from the raw text, the AQE model is used to evaluate the quality of the arguments. The models that achieved the best validation scores on the training data were all fine-tuned versions of BERT. The best performing model was a RoBERTA model (Liu et al., 2019), trained on roughly 1,100 samples contained within the GAQCorpus (Lauscher et al., 2020).

Before using the data for fine-tuning, we preprocess it to make the arguments more closely aligned to the arguments found in scientific literature. The original dataset spans three different domains: *debate* forums, *answer* forums and *review* forums. We exclude the data taken from the *review* forum from our training data, as these samples are the least argumentative and most subjective in terms of content. There is also a binary feature included in the data that determines whether the annotators deemed the sample argumentative or not - we remove all samples where there is not a unanimous agreement that the sample is argumentative.

Furthermore, we make slight modifications to the remaining samples in our training set to increase syntactical similarity with the scientific arguments. Rhetorical questions are frequently used in the forum data, which is not found in any scientific content. There is also use of very short sentences (five words or less), which is practically non-existent in scientific writing. Therefore, we remove any sentences that fit into either of these categories, as well as converting any extraneous punctuation (e.g. exclamation marks) into full stops. An example of the preprocessing is presented in Table 5.

Instead of using an average of the three metrics which are contained within the GAQCorpus, we only make use of the *cogency* rating. This is due to the relevance of cogency to scientific argumentation, and the relative irrelevance of the other two metrics (*reasonableness* and *effectiveness*) within the context of scientific literature. To illustrate this, the definition for *cogency* used by Lauscher et al. (2020) to guide the annotators was: '[cogency] relates to the logical aspects of [argument quality], for instance, whether an argument's premises are acceptable (local acceptability) or whether they can be seen as relevant for the conclusion (local relevance)'.

3.4 Full Argumentation-Based System

The final system in its entirety takes the output from the argument identification model and uses it to perform a transformation of the input text data. The transformation is a linearisation (Stede and Sauermann, 2008) of the extracted argumentative components, so that a string can be used as input to the AQE model, reflecting the training data. This string consists of the claim sentence(s) followed by the premises. This is chosen as it most closely resembles the format of the arguments in the non-scientific training corpus (GAQCorpus).

In order to make the final classification into 'legitimate' or 'fraudulent', it is necessary to establish a threshold (T) which the quality score can be compared to. There are five quality classes in total. In order to create a system that is less likely to return false positives (classify fraudulent articles as legitimate), we only consider an argument as legitimate if it is in the highest two classes of quality. If it is in the bottom three classes of quality we classify the sample as fraudulent.

Our decision to choose a threshold weighted towards the classification of samples as fraudulent was done with the practical purpose of the finished system in mind, as well empirical validation. Relative to false negatives (classifying fraudulent documents as legitimate), false positives (flagging legitimate documents as fraudulent) are less detrimental

to a system which is built to assist in the detection of academic fraud.

4 Results and Discussion

In Section 4.1 we present a comparison of our novel AM model with two existing alternatives. Then, we outline the performance on the overall fraud prediction task of three novel baselines (Section 4.2) and the full argumentation-based model (Section 4.3).

4.1 Scientific Argument Mining

In order to evaluate our AM model, described in Section 3.2, we use two of the best performing models developed by Fergadis et al. (2021) as baselines. All three models are trained on the SciARK dataset and tested on the AbstRCT dataset.

The results in Table 2 show that the overall F_1 score was best when using our novel *scibert-transformer*. As previously mentioned, we infer that the transformer's capability to efficiently handle long-range dependencies - compared to the BiLSTM used in the second best model - is advantageous for this task.

4.2 Baselines for Fraudulent Paper Detection

So that we could provide insightful benchmarks and ablations for the fraudulent paper detection task in its entirety, we compared our argumentative approach with five baselines. This was necessary due to the novelty of our research, and the consequent absence of existing systems in the literature that performed a comparable function. Three of these benchmark utilise BERT-style modes, mirroring our main method. The other two use autoregresssive LLMs, Mistral (Mistral-7B-Instruct-v0.2) (Jiang et al., 2023) and Mixtral (Mixtral-8x7B-Instruct-v0.1) (Jiang et al., 2024).

Our first model, *SciBERT direct inference*, was designed to infer legitimacy directly. Instead of training the model to determine the quality of the reasoning in a sample as an intermediary step, we trained it with legitimate and fraudulent samples directly. For our fraudulent training samples, we collected a separate training set of 556 fraudulent samples from the Retraction Watch database. These samples were taken from a diverse range of domains, in order to reflect the diversity of domains present in the SciARK corpus which we used for our training set of legitimate samples. Thus we limited the possibility that the performance of the model was influenced by subject matter, and instead learnt the 'legitimacy' and 'fraudulence' features present in the respective samples.

We tested various base LLMs and identified that fine-tuning a SciBERT model resulted in the best performance. However, its performance was still lacking. We propose that a significant reason for this was the limited number of negative samples contained in our dataset, which restricted the model's ability to capture a comprehensive representation of the sample space.

The second baseline we investigated, *full text quality*, simply skipped the argument identification stage, and used the original, unmodified abstracts as input to the AQE model. Similarly, for our third baseline we first summarised the abstracts using a LLM, Mixtral, as a form of feature extraction, and once again used that as input to the AQE model.

The summarisation technique produced better results than using the full text. However, it was still relatively ineffective compared to the full argumentation-based approach. Both the summary and the argumentative content of a piece of text are comparable features, but our belief is that the argumentation-based approach provides a more faithful representation of the quality of reasoning in a passage of text.

For our fourth and fifth baselines we leveraged the task-agnostic, general capabilities of state-of-the-art LLMs. For both the Mistral and Mixtral models, we use zero-shot chain-of-thought (COT) (Wei et al., 2022; Zhang and Parkes, 2023) prompting. The full prompt can be seen in Appendix C. All outputs were generated with greedy sampling (equvalent to setting temperature to 0). As can be observed in Table 3, the Mixtral model performed the best out of all the baselines. This is especially notable in light of the zero-shot setting in which the experiments were conducted. This introduces the prospect that the use of even larger LLMs could present further gains in performance. This conjecture also applies to the use of larger LLMs as the components of our argumentative system.

4.3 Full Argumentation-Based Model for Fraudulent Paper Detection

The proposed method demonstrates a considerable improvement over the baselines, as illustrated by Tables 3 and 4. By comparing these results to the five distinct baselines we developed, we highlight that both the AM and AQE components contributed to the improved performance of the system.

The favourable comparison with the *direct in-*

Model	Evidence			Claim			Average
	P	R	F_1	P	R	F_1	F_1
SciBERT-Only	0.778	**0.728**	**0.752**	**0.808**	0.359	0.497	0.625
SciBERT-Bilstm	0.815	0.688	0.747	0.874	0.458	0.601	0.674
SciBERT-Transformer	**0.878**	0.57	0.693	0.858	**0.557**	**0.675**	**0.684**

Table 2: Performance metrics for the Argument Mining models described in Section 3.2. We compare our novel *Scibert-Transformer* model with the previous models introduced in Fergadis et al. (2021). All models are trained on the SciARK corpus and evaluated on the AbstRCT corpus (Mayer et al., 2020).

Model	Accuracy	Precision	Recall	F_1
SciBERT Direct Inference	0.543	1	0.109	0.197
Full Text Quality	0.461	1	0.012	0.024
Summary Quality	0.463	0.676	0.23	0.343
Mistral COT	0.622	0.021	1	0.041
Mixtral COT	0.637	0.429	0.537	0.476
Argumentation-Based Model	0.761	0.708	0.648	**0.677**

Table 3: Performance metrics for the baseline models described in Section 4.2 and our novel methodology *Argumentation-Based Model*.

	Predicted Fraudulent	**Predicted Legitimate**
Actual Fraudulent	272	148
Actual Legitimate	112	557

Table 4: Argumentation-Based Model's performance on 1,089 legitimate and fraudulent abstracts.

ference model validates that the development of a more intricate approach is appropriate and worthwhile. Likewise, the comparisons with the *full text quality* and *summary quality* models, give credence to the hypothesis that the evaluation of arguments, rather than full texts or summaries, is beneficial to the system's overall performance.

As well as performing better, relative to directly training a model on legitimate and fraudulent samples, the methodology developed for evaluating argument quality provides a more flexible and generalisable approach. This is due to the noise inherently present in a training set that consists solely of scientific abstracts. There is no clear way to disentangle useful properties of the data - namely, legitimacy of the article - from noisy properties such as subject matter and syntactical idiosyncrasies. This would not present such a significant obstacle if the available data for training were large and diverse enough, but as this is far from being the case it must be taken into consideration.

We also gain valuable insights by comparing the results of the fully argumentative model with the two baselines that provide different inputs to the AQE model. We observe that using the arguments contained within the abstracts as representative features is more effective than using the full text, or using summaries as a feature. There are a number of possible explanations for this finding. One plausible suggestion is that mining the arguments contained within a scientific abstract is a good way of extracting a representation of the reasoning contained therein, and, furthermore, the quality of this reasoning is indicative of the article's legitimacy. This would confirm our initial hypothesis.

Another factor to consider is the architecture of the AQE model we have developed. The training data that we used from the GAQCorpus (Lauscher et al., 2020) is composed of samples that have been evaluated for their argumentative quality. The metric we choose to consider (*cogency*), is something that is relevant to argumentation, but not so much for summaries.

4.4 Qualitative Error Analysis

In this section we provide examples and analysis of mined arguments that were falsely classified by the AQE model as legitimate and fraudulent respectively.

Incorrectly classified as legitimate

> Compared to GES-1 cells, the expressions of miR-214, -catenin and survivin in MKN-28 cells were upregulated, along with downregulation of GSK-3 expression. After the transfection of miR-214 inhibitor and/or pSicoR-GSK-3, GSK-3 expression was induced in MKN-28 cells while -catenin and survivin expressions were inhibited, along with the increase of cell apoptosis.

Incorrectly classified as fraudulent

> At 6 months after the end of RT, global HRQOL was higher in the TPF arm than in the PF arm, but the low compliance does not allow to draw definitive conclusions. Swallowing and coughing problems decreased more in the TPF arm than in the PF arm at the end of cycle 2, but to a limited extent.

Both examples demonstrate the complexity of the task. The first example, which is taken from a fraudulent abstract, almost entirely consists of technical terminology. While the data used to train SciBERT consists of scientific text, medical literature contains a large amount of domain-specific language, which is sometimes exclusively used by the community working on the specific problem. Besides this, the content of the argument also seems to be sound. This is inevitable due to the scale of papers produced fraudulently, so must be taken into consideration in the context of our solution. It is essential that any fully-fledged system takes into account other factors that may indicate fraudulence, such as the presence of irregular images or data. In this case argument quality may be considered as one out of many features.

For the second example, taken from a legitimate abstract, one might note some unorthodox grammatical constructions as reasons for the fraudulent classification. For example, 'the low compliance does not allow to draw definitive conclusions' contains a slight grammatical error ('allow *one* to draw' would be a more sound construction). While this is entirely reasonable to expect in a legitimate manuscript, it may lead to a fraudulent classification as it is less commonly found in legitimate papers compared to those produced by a paper mill.

It is important to note that the above are human interpretations of the data, and may not be faithful to the true underlying processes carried out by the model. Future work could be undertaken to apply established methods from the explainable AI literature. Furthermore, there are a diverse range of incorrect classified samples, with the examples chosen being representative of one type. The reasons for these errors is likely to vary across inputs.

5 Conclusion and Future Work

The aim of this study was to determine whether it is possible to automate the detection of fraudulently produced scientific publications. To achieve this, we introduced and implemented a novel framework, building upon existing architectures from the argument mining literature.

We developed state-of-the-art methods in the fields of scientific argument mining and argument quality evaluation. By using both models in conjunction, we created a method for evaluating the quality of reasoning in scientific articles.

We compared this approach to three baselines, observing favourable comparisons in each case. By using a variety of baselines we were able to demonstrate that both the argument mining and quality evaluation components positively contributed to the overall performance of the system.

Although the initial results were promising, there are a number of potential developments that would merit further exploration. For instance, the quality evaluation component would benefit from being trained on arguments that have been taken directly from scientific papers. Furthermore, introducing methods to assess the quality of the individual premises and claims, in addition to the overall argument, could benefit performance.

As mentioned previously, there are various advantages to restricting our inputs to abstracts. However, analysing entire articles also has benefits. State-of-the-art LLMs have made this plausible. An evident direction for future work would be the utilisation of larger, more sophisticated language models, while keeping in mind the trade-off between performance and computational cost.

Finally, widening the range of modalities that are included in the analysis would lead to a more comprehensive system. It is standard practice in scientific articles to include arguments that consist of more than just textual components. Images, graphs, citations and tabular data are all commonly incorporated as sources of evidence. An ideal system would be able to assess the quality and relevance of all these forms of data with respect to the arguments contained within an article.

Limitations

The system has not yet been sufficiently scrutinised to confidently assert that it could be effectively deployed in a real-world setting. Due to the sensitivity of the domain, it is important that the system undergoes extensive testing and is validated by individuals with expertise in fraudulent article detection, before it can be effectively and safely deployed.

Due to the inefficiency of the existing human-centric process of detecting fraudulent articles, there is a relatively small dataset of retracted articles to use for evaluation. Therefore some caution should be exercised in making assumptions about the totality of articles that have been published using illegitimate means, as there may be bias present in the subset of articles that have already been detected.

Ethics Statement

Academic Fraud is an important and sensitive issue and any attempts to automate its detection must be approached with some degree of caution. It is essential that any such tool is used in conjunction with human experts, and is not used in isolation to make decisions.

Acknowledgements

This research was partially supported by ERC under the EU's Horizon 2020 research and innovation programme (grant agreement No. 101020934), by J.P. Morgan and the Royal Academy of Engineering under the Research Chairs and Senior Research Fellowships scheme and by UKRI through the CDT in Safe and Trusted Artificial Intelligence (Grant No. EP/S023356/1) and through the INDICATE project (Grant No. EP/Y017749/1). We thank Retraction Watch and Grammarly for providing access to their respective datasets.

References

Anna Abalkina. 2021a. Detecting a network of hijacked journals by its archive. *Scientometrics*, 126:7123 – 7148.

Anna Abalkina. 2021b. Publication and collaboration anomalies in academic papers originating from a paper mill: evidence from a russia-based paper mill.

Pablo Accuosto, Mariana L. Neves, and Horacio Saggion. 2021. Argumentation mining in scientific literature: From computational linguistics to biomedicine. In *BIR@ECIR*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In *Conference on Empirical Methods in Natural Language Processing*.

Frank Biermann, Norichika Kanie, and Rakhyun E. Kim. 2017. Global governance by goal-setting: the novel approach of the un sustainable development goals. *Current Opinion in Environmental Sustainability*, 26:26–31.

Mark J. Bolland, Andrew Grey, and Alison Avenell. 2022. Citation of retracted publications: A challenging problem. *Accountability in Research*, 29(1):18–25. PMID: 33557605.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, S. Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen A. Creel, Jared Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas F. Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, O. Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Benjamin Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, J. F. Nyarko, Giray Ogut, Laurel J. Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Robert Reich, Hongyu Ren, Frieda Rong, Yusuf H. Roohani, Camilo Ruiz, Jack Ryan, Christopher R'e, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishna Parasuram Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei A. Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2021. On the opportunities and risks of foundation models. *ArXiv*, abs/2108.07258.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec

Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *NIPS*.

Declan Butler. 2018. Researchers have finally created a tool to spot duplicated images across thousands of papers. *Nature*, 555:18–18.

Cristina Candal-Pedreira, Alberto Ruano-Raviña, and Mónica Pérez-Ríos. 2021. Should the european union have an office of research integrity? *European journal of internal medicine*.

Joyita Chakraborty, Dinesh K. Pradhan, and Subrata Nandi. 2021. On the identification and analysis of citation pattern irregularities among journals. *Expert Systems*, 38(4):e12561.

Oana Cocarascu, Elena Cabrio, Serena Villata, and Francesca Toni. 2020. Dataset independent baselines for relation prediction in argument mining. In *Comma*.

Clive Cookson. 2023. Study reveals scale of 'science scam' in academic publishing. *The Financial Times*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805.

The Economist. 2023. There is a worrying amount of fraud in medical research. *The Economist*.

Holly Else and Richard van Noorden. 2021. The fight against fake-paper factories that churn out sham science. *Nature*, 591:516 – 519.

Daniele Fanelli, Julie Wong, and David Moher. 2022. What difference might retractions make? an estimate of the potential epistemic cost of retractions on meta-analyses. *Accountability in Research*, 29(7):442–459. PMID: 34196235.

Aris Fergadis, Dimitris Pappas, Antonia Karamolegkou, and Haris Papageorgiou. 2021. Argumentation mining in scientific literature for sustainable development. In *Proceedings of the 8th Workshop on Argument Mining*, pages 100–111, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Craig A Garmendia, Liliana Nassar Gorra, Ana Lucia Rodriguez, Mary Jo Trepka, Emir Veledar, and Purnima Madhivanan. 2019. Evaluation of the inclusion of studies identified by the fda as having falsified data in the results of meta-analyses: the example of the apixaban trials. *JAMA Internal Medicine*, 179(4):582–584.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks : the official journal of the International Neural Network Society*, 18 5-6:602–10.

Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. 2019. A large-scale dataset for argument quality ranking: Construction and analysis. In *AAAI Conference on Artificial Intelligence*.

Ivan Habernal and Iryna Gurevych. 2016. Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional lstm. In *Annual Meeting of the Association for Computational Linguistics*.

Robin Haunschild and Lutz Bornmann. 2021. Can tweets be used to detect problems early with scientific papers? a case study of three retracted covid-19/sars-cov-2 papers. *Scientometrics*, 126:5181 – 5199.

Yassine Himeur, Somaya Ali Al-Maadeed, Noor Almadeed, Khalid Abualsaud, Amr Mohamed, Tamer M. S. Khattab, and Omar Elharrouss. 2022. Deep visual social distancing monitoring to combat covid-19: A comprehensive survey. *Sustainable Cities and Society*, 85:104064 – 104064.

Joanne Horton, Dhanya Krishna Kumar, and Anthony Wood. 2020. Detecting academic fraud using benford law: The case of professor james hunton. *Research Policy*, 49(8):104084.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts.

Omkar Joshi, Priya Pitre, and Yashodhara V. Haribhakta. 2023. Arganalysis35k : A large-scale dataset for argument quality analysis. In *Annual Meeting of the Association for Computational Linguistics*.

Soo Young Kim, Hyun Jung Yi, Hye-Min Cho, and Sun Huh. 2019. How many retracted articles indexed in koreamed were cited 1 year after retraction notification. *Science Editing*.

Nick Kinney, Araba Wubah, Miguel Roig, and Harold R. Garner. 2021. Estimating the prevalence of text overlap in biomedical conference abstracts. *Research Integrity and Peer Review*, 6.

Anne Lauscher, Lily Ng, Courtney Napoles, and Joel Tetreault. 2020. Rhetoric, logic, and dialectic: Advancing theory-based argument quality assessment in natural language processing. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4563–4574, Barcelona, Spain (Online). International Committee on Computational Linguistics.

John Lawrence and Chris Reed. 2020. Argument mining: A survey. *Computational Linguistics*, Just Accepted:1–55.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Adam Marcus and Ivan Oransky. 2023. The retraction watch database. *New York: The Center for Scientific Integrity*.

Andreas F. Mavrogenis and Marius M. Scarlat. 2023. Quality peer review is mandatory for scientific journals: ethical constraints, computers, and progress of communication with the reviewers of international orthopaedics. *International Orthopaedics*, 47:605–609.

Tobias Mayer, Elena Cabrio, and Serena Villata. 2020. Transformer-based argument mining for healthcare applications. In *ECAI 2020 - 24th European Conference on Artificial Intelligence*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, pages 2108–2115. IOS Press.

Andreas Peldszus and Manfred Stede. 2013. From argument diagrams to argumentation mining in texts: A survey. *Int. J. Cogn. Informatics Nat. Intell.*, 7:1–31.

N. Pooranam, Priya P N Sushma, Sai Sruthi, and Dhanya K Sri. 2021. A safety measuring tool to maintain social distancing on covid-19 using deep learning approach. *Journal of Physics: Conference Series*, 1916.

Ramon Ruiz-Dolz, José Alemany, Stella Heras Barberá, and Ana García-Fornes. 2020. Transformer-based models for automatic identification of argument relations: A cross-domain evaluation. *IEEE Intelligent Systems*, 36:62–70.

BA Sabel, E Knaack, G Gigerenzer, and M Bilc. 2023. Fake publications in biomedical science: Red-flagging method indicates mass production.

Bernhard A. Sabel and Roland Seifert. 2021. How criminal science publishing gangs damage the genesis of knowledge and technology—a call to action to restore trust. *Naunyn-Schmiedeberg's Archives of Pharmacology*, 394:2147 – 2151.

Edwin Simpson and Iryna Gurevych. 2018. Finding convincing arguments using scalable bayesian preference learning. *Transactions of the Association for Computational Linguistics*, 6:357–371.

Christian Stab and Iryna Gurevych. 2016. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43:619–659.

Manfred Stede and Antje Sauermann. 2008. Linearization of arguments in commentary text.

Luke Thorburn and Ariel Kruger. 2022. Optimizing language models for argumentative reasoning. In *ArgML@COMMA*.

Assaf Toledo, Shai Gretz, Edo Cohen-Karlik, Roni Friedman, Elad Venezian, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim. 2019. Automatic argument quality assessment - new datasets and methods. In *Conference on Empirical Methods in Natural Language Processing*.

Richard van Noorden. 2021. Hundreds of gibberish papers still lurk in the scientific literature. *Nature*, 594:160 – 161.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187, Valencia, Spain. Association for Computational Linguistics.

Elizabeth Wager, Sabine Kleinert, Michele Garfinkel, Volker Bahr, Ksenija Badarić, Michael J. G. Farthing, Chris Graf, Zoë Hammatt, Lyn Horn, Susan King, Debra Parrish, Bernd Pulverer, Paul Taylor, and Gerrit van Meer. 2017. Cooperation & liaison between universities & editors (clue): recommendations on best practice. *Research Integrity and Peer Review*, 6.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS 2022*.

Hugh Zhang and David C. Parkes. 2023. Chain-of-thought reasoning is a policy improvement operator. *CoRR*, abs/2309.08589.

Yue Zhao, Ajay Anand, and Gaurav Sharma. 2021. Reviewer recommendations using document vector embeddings and a publisher database: Implementation and evaluation. *IEEE Access*, 10:21798–21811.

A Hardware Configuration

All models presented in this paper were trained on a computing cluster containing the following GPUs:

- Nvidia Tesla A30 with 24GB RAM
- Nvidia Tesla T4 with 16GB RAM
- Nvidia GeForce GTX Titan Xp with 12GB RAM

B Code and Data

The following repository contains all relevant code and data: https://github.com/GIFRN/Scientific-Fraud-Detection/tree/main.

We also present a detailed breakdown, including hyperparameters, of the AQE model that we use for all experiments. This is because we include the trained model already in the repository, where as for the AM modle we include the training script and training data. Hyperparameters were optimised by means of extensive grid search.

B.1 Argument Quality Evaluation model

The same architecture is used for the results reported in Tables 3 and 4, and for both of the 'Quality' models in Table 3.

Input dimensions
Max token length: 512

RoBERTa Model
Number of training epochs: 5
Number of folds: 5
Dropout rate: 0.2
Batch size: 8

AdamW Optimizer
Learning Rate: 1×10^{-5}
Weight Decay: 1×10^{-2}

C Prompt

The following is the prompt used for both the Mistral and Mixtral baselines described in Section 4.2:

```
"Please return whether the following
scientific abstract is fraudulent or
legitimate. A fraudulent abstract
is one that has been produced
by a paper mill or has not
undergone a proper peer review process.
Please think through your answer
step by step before returning
your final answer.
Present your final answer as
'STATUS: Legitimate'
if the abstract is legitimate or
'STATUS: Fraudulent'
if the abstract
is fraudulent.\nAbstract: " + abstract
```

D Argument Quality Evaluation Preprocessing

In Table 5 we show two examples of samples taken from the GAQCorpus, used for training the AQE model. We show the samples before and after preprocessing, and include the rating.

Original Text	**Preprocessed Text**	**Cogency Rating**
Wow, that's tough. Is your assignment to argue that retribution is socially cohesive, or did you come up with this yourself? This is what I can think of: Retributive justice is based on the idea that criminals should be punished for undermining social "harmony" or "balance." Therefore, we can't dole out retributive punishment before we first have a definition of what is social harmony or balance. Social harmony and balance can be defined through laws, custom, or religion. The PROCESS of defining social harmony and balance is socailly cohesive because, to some extent, we must generally agree on what the definition of social order is. In other words, the PROCESS of accepting and agreeing on laws/customs/religion is socially cohesive. The rehabilitation theory of punishment is probably the most socially cohesive because it is based on rehabiliting the criminal so that he can successfully reenter society.	This is what I can think of: Retributive justice is based on the idea that criminals should be punished for undermining social "harmony" or "balance. " Therefore, we can't dole out retributive punishment before we first have a definition of what is social harmony or balance. Social harmony and balance can be defined through laws, custom, or religion. The PROCESS of defining social harmony and balance is socailly cohesive because, to some extent, we must generally agree on what the definition of social order is. In other words, the PROCESS of accepting and agreeing on laws/customs/religion is socially cohesive. The rehabilitation theory of punishment is probably the most socially cohesive because it is based on rehabiliting the criminal so that he can successfully reenter society.	4
Am I reading this right? A pot head is waging war on a meth head?! I never thought this sort of reasoning would make it farther than the 'idea' having moment during a toking. You got more tar in your head than you do brains. And yes, Meth addicts DO need jail time. Every one of them made a clear and concious decision to pick up that hot rail, needle or pipe to smoke crystal when they started and the time it took to become 'uncontrolably' addicted, so they accepted the risks. Besides, MOST users sell, they keep cutting some out every time it exchanges hands, so lock 'em up!! And I can't belive someone is trying to defend the 'poor' addicts. What and idiot you are Cripple play!!!	I never thought this sort of reasoning would make it farther than the 'idea' having moment during a toking. You got more tar in your head than you do brains. And yes, Meth addicts DO need jail time. Every one of them made a clear and concious decision to pick up that hot rail, needle or pipe to smoke crystal when they started and the time it took to become 'uncontrolably' addicted, so they accepted the risks. Besides, MOST users sell, they keep cutting some out every time it exchanges hands, so lock 'em up. And I can't belive someone is trying to defend the 'poor' addicts. What and idiot you are Cripple play.	2

Table 5: Illustrative examples from the GAQCorpus, before and after preprocessing