We would like to thank the editor and the reviewers for their constructive comments. See below our point-by-point response.

Reviewer 1

1. The description, under 'Data', of how cases were selected needs clarification in several ways (lines 165-182).
First, when it is said that Article 3 'Prohibits torture', are we to understand that the study does not cover the other prohibitions contained in Article 3 (such as the prohibition on inhuman treatment)? Precision is important in legal writing and it is important here.

The study does cover the other prohibitions contained in Article 3 (such as inhuman and degrading treatment). Accordingly, we have added the full description of the Articles in Table 1.

2. Second, the number of cases seems much lower than what one would expect, and much lower than what a rudimentary search of the HUDOC database generates. For example, a basic HUDOC search of Article 6, in English, generates over 10,000 cases. Of those, HUDOC indicates there at least 8,000 violation cases and at least 900 non-violation cases. If the methodology at lines 172-177 is replicable, one would expect that this study would have included at least 1,800 cases in its study of Article 6 (being all the cases in the smaller class plus a randomly selected equal number of cases from the larger class). And yet here the number studied is just 80. It is hard to discern the reasons for the discrepancy between the draft article and the results of a basic HUDOC search.

At the beginning of our study, we planned to manually develop the data set using experts in the School of Law of the University of Sheffield. We quickly realised that this process was very slow and thus infeasible. Then, we decided to automate the process by devising a “reasonable” common structure in the format of the case reports. This includes the main parts of “Procedure”, “The Facts”, “The Law” and “Operative Provisions” in that order. We strictly filtered out cases that failed to match more than one of these main sections. There are many cases that a different wording is used in the title of a section which makes it difficult to be captured automatically. In addition, we strictly filtered out comments made by the Court keeping only those comments made by the two parties. That sometimes resulted into empty sections. Finally, many case reports retrieved were actually in French even if the selected language was set to English. For these reasons, our dataset contains a smaller number of cases. However, the results obtained are significantly different compared to the random baseline, i.e. 50% accuracy (t-test, $p<0.001$). Note that we scraped the Hudoc website and matched cases with regular expressions without having access to the actual database (access here means to be able to retrieve data automatically using some sort of database query language such as SQL and not through the website interface). We believe that our results can be seen as a proof of concept and by given access to the actual database we could perform a study that covers all the available cases and articles. That would be an interesting avenue for future work and/or a research grant proposal.

3. Third, the reasons for choosing articles 3, 6, and 8 could be substantiated a bit more. Surely *all* of the ECHR rights may be regarded as "important human rights that correspond to a variety of interests" (lines 167-168)? Why focus on these three?

These Articles seemed to us to provide the most data we could automatically scrape. We have revised the text accordingly.

4. The article claims that "there is a strong correlation between the actual facts of a case and the decisions made by judges" (lines 264-265). I have serious concerns about whether or not this conclusion is substantiated by the data which preceded the Discussion. First, it is unclear what the authors mean by "the *actual* facts" (line 265). Are the "actual" facts somehow different from "the facts"?

No, they are not. We have revised the text accordingly.

5. Second, it is not clear what the authors mean when they say "information available to the judges before they make any comments or decisions" (line 180). Are the authors implying that the judgments contain all the information available to the judges before they made their decision? If so, this would seem to be a misunderstanding of how courts work (surely, at the very least, one would want to look at the full written arguments of the parties, rather than simply the summaries of those written arguments that are contained in judgments?).

Our wording was somewhat sloppy. All we meant to say was that the models do not have information pertaining to the operative provisions. We have revised the text accordingly.

6. Third, it seems to me that this study proves, at best, that there is a correlation between the facts *as described in the judgement* and the result of a case. There is a difference between "the actual facts of a case" and "the facts as they are described in the judgment of case". The article does not acknowledge this difference at all. This is a problem. I'm afraid that the authors seem to be under the impression that the facts section of a judgment is an objective scientifically-established recitation of the facts. Unless the authors are aware of ECtHR practice that I am unaware of, this seems dangerously naive. On my understanding, the judgments of the ECtHR are prepared by the judges, their assistants, and the Court Registry. In any court anywhere around the world, including the ECtHR, it would not be unusual in the slightest for the judges, the assistants, or the registry, to frame the facts in light of their full understanding of the case (which would include their view on whether or not there is a violation). Facts sections of judgments are not peer-reviewed scientific papers. They are subjective summaries of the facts, including what the authors think is relevant and what they think is irrelevant. If a judge/judicial assistant/registrar is of the preliminary view that

a violation is likely, it would not be at all unusual for them to frame the facts differently than how those same facts would be framed if they were of the view that a violation was unlikely. This raises a problem: the article involves the authors taking the facts section of a delivered judgment, and then predicting whether or not that same judgment will result in a violation or not. This may be useful, but it does not seem to be the same as "predicting judicial decisions...using only the textual information available to the judges before they make any comments or decisions about a specific case..." (lines 312-313). The model does not seem to provide any capability for ex ante prediction -- i.e. it does not allow the result of a judicial decision to be predicted until the facts section of the judgment can be analysed (and the facts section of the judgment cannot be analysed until the judgment is handed down). Surely this limits its utility? Perhaps I misunderstand the mathematical value of the study; perhaps I misunderstand the internal workings of the European Court. But even if I am wrong on the maths or on the workings of the Court, the article needs to be considerably clearer about what it is predicting and about the nature of how judgments are written and prepared. Without that, it is hard to attach too much significance to its findings, I'm afraid.

We should have made our argument clearer here. We made revisions to the text accordingly, stressing the following points. First, the ECtHR has only limited fact-finding powers, which implies that, in the vast majority of cases, it will defer, when summarizing the facts, to the judgments of domestic courts that have already heard and dismissed the applicants' complaint. While these can also reflect assumptions about relevance, they also reflect understandings of the facts that have been validated by more than one decision-maker. Second, the Court cannot openly acknowledge any kind of bias on its part. This means that, on their face, summaries of facts have to be at least framed in as neutral a way as possible. Furthermore, a random reading of ECtHR cases indicates that, in the vast majority of cases, parties do not seem to dispute the facts themselves, but merely their legal significance (i.e. whether a violation took place or not, given those facts). Third, it is important to note that the data used by the model are to do with 'the facts of the case are these are described in the relevant section of the judgment', as the reviewer correctly suggested. We have revised all pertinent formulations accordingly. Fourth, for our argument to get off the ground all we need is that the text of this section performs differently from the text of other sections. This much has been established by our model. Fifth, the reviewer is right that we should deflate our claim: the model is only a (crude) proxy to different kinds of considerations, and not a perfect representative of these considerations. We have revised all pertinent formulations accordingly. Sixth, this of course leaves open the possibility, noted by the reviewer, that this section is indeed formulated in a way that reflects judges' understanding of the case, which includes various judgments relating to relevance/irrelevance and, potentially, to biases related to how the case should be decided. We have acknowledged this openly, revising all pertinent formulations. Seventh, and final, point: insofar as the 'facts' section of the case is a (crude) proxy, it is an open question whether it could provide a basis for ex ante predictions of judgments. We do not really see any reason why it could not, since it — at the very least — proves the concept that, on the basis of chunks of particular textual information that differ on their face, it can do a relatively good job at predicting outcomes. So the model could have practical utility in this respect.

7. Fourth, the authors claim that their study amounts to support for legal realism over legal formalism (line 322). This may be so, but a much more sophisticated account (than that at lines 28-37) of the debate about realism and formalism would be needed to draw much of a conclusion here.

We have (a) moved the relevant points from the introduction to the discussion part and (b) deflated our claims accordingly, with all the usual caveats. Different (sub)sections of a judgment are not to be understood as more than crude proxies (but they are all that we have at this point). Again, we believe that the important thing is that the model, given the data, differentiates clearly between (sub)sections of a judgment and that is a significant result in itself.

Reviewer 2

8. The research question is clearly defined: it is possible to use text processing and machine learning to predict whether, given a case, there has been a violation of an article in the convention of human rights. The research question is certainly relevant, and the results are interesting for the natural language processing and machine learning community, but it's unclear how these findings are useful for the law and human rights community, since nothing is mentioned in the paper with regards to this question, for example, 'it would be useful to apply this kind of classifier as a tagging tool for highlighting cases in which violation of human rights are likely to be true? perhaps as a prioritizing or filtering means?'

From a more "applied" perspective, we mention in the abstract that "This can be useful, for both lawyers and judges, as an assisting tool to rapidly identify cases and extract patterns which lead to certain decisions.". In the revised version we have added that in the introduction as well. Moreover, insofar as different sections of the judgment can be understood as (crude) proxies of the relevance of different kinds of considerations to judicial decision-making, the analysis provides a first step that could be later further tested with text coming from lawyers' briefs/applications or domestic judgments. The hard part is to have access to that data (and that is why we focused on the ECtHR's judgments).

9. Topic models: in spectral clustering, the number of topics is input parameter, but nothing is mentioned about how the value of this parameter, in this case 30 topics, was chosen.

We tuned this parameter using the same strategy followed to tune the SVM parameters. We have added the description of how we set the value of this parameter in the Classification Model subsection.

10. I was expecting to see experiments using both set of features, bow and topics, but results are only reported for experiments using one set of features at a time. Why are there no experiments that combine both features sets? If the performance was lower than when using the individual features sets, the outcome is still useful for the community and should be reported.

We performed experiments combining both sets of features (N-grams Circumstances + Topics) yielding slightly better performance for articles 6 and 8 while performance was slightly lower for article 3. That is 0.75 (0.10), 0.84 (0.11) and 0.78 (0.06). We have updated Table 2 accordingly.

11. Accuracy is reported as the evaluation metric used to measure performance, but nothing is said about how accuracy is calculated, the formula or an explanation would be helpful. I'm assuming accuracy should be understood to mean: the proportion of

true outcomes (true positives and true negatives) among the total number of cases. Please, clarify this.

That is true. We have added the equation in section “Results and Discussion”.

12. In the discussion section, the authors gave examples how topics aligned with the theme of some of the cases, but it’s hard to understand if those examples are from the dataset used or from another dataset. I figured it out by exploring the dataset itself. A line clarifying this would be helpful.

The examples of cases we use in the discussion are from the dataset used in the study. We clarify that in the revised version of the paper.

13. Comment something about the cases that were wrongly classify, for example, is there any commonality between the wrongly classified instances? What does it mean for the law community to have more than 20% of its cases wrongly classified?

On the other hand, cases have been misclassified mainly because their textual information is similar to cases in the opposite class. We observed a number of cases where there is a violation having a very similar feature vector to cases that there is no violation and vice versa. We have added that comment in the Discussion subsection (Results section).

14. The topic of the paper is very interesting and the paper is easy to follow, but I would like to read about how this results can be use by the law community.

See point 8.