Trawling through the Rubbish: Data Mining of the Scientific Literature on Marine Plastic Pollution

Qixiang Fang, Mehran Moazeni, Sebastian Mildiner Moraga
Department of Methodology & Statistics, Utrecht University
{q.fang, m.moazeni, s.mildinermoraga}@uu.nl

April 6, 2021

Information about Project and Team members

- Title: Trawling through the rubbish: Data mining of the scientific literature on marine plastic pollution
- Team leader: Qixiang Fang
- Representative of the problem: Erik van Sebille
- Team members: Qixiang Fang, Mehran Moazeni, & Sebastian Mildiner Moraga

1 Introduction

1.1 Challenge Context

Loads of litter every year find their way in seas and oceans (Sebille et al., 2015). Plastic derivatives of diverse characteristics have been documented in the full range of marine environments, from shorelines to open waters (Barnes, Galgani, Thompson, & Barlaz, 2009). They have been identified on surface waters and sea floor alike (Schlining et al., 2013), and even in the Mariana trench (Chiba et al., 2018), the deepest point of the oceans. Plastic debris is known to interfere with the environmental flora and fauna in many different ways (Wayman & Niemann, 2021), and yet little is known about its distribution and degradation cycles (Wayman & Niemann, 2021). For historic and economic reasons, our knowledge of its distribution is biased towards specific regions. For example, areas with higher commercial and demographic impact such as the Western North Atlantic Ocean (Morét-Ferguson et al., 2010) and the Eastern North Pacific Ocean (Goldstein, Titmus, & Ford, 2013; Law et al., 2014) have been subject to *in situ* surveys (e.g. neston net trawls) aimed at determining the characteristics and composition of the plastic debris. Meanwhile, other regions have received less attention, resulting in a significantly scarcer wealth of knowledge for subtropical regions and the southern hemisphere as a whole (Cózar et al., 2014). The same uneven pattern exists in the quality of the information surveyed: most of the measurements available come from surface and coastal debris, even though as much as 90% of the plastic in the sea is hypothesised to reside in the deep-sea sediments (Booth et al., 2017). Kaandorp, Dijkstra, and van Sebille (2020) argue that combining the strengths of numerical methods and *in situ* measurements provides with an effective framework to expand our understanding on the amount and distribution of plastic debris on different compartments of the sea.

The main bottleneck to modelling the temporal distribution of plastics in the sea is the absence of a curated database that retrieves and aggregates the information from collected with *in situ* surveys. Researchers report these measurements concerning the tempo distribution, characteristics, and concentrations in their publications. As a result, it is difficult and extremely time-consuming to identify most (if not all) these publications, let alone download, read, and look for relevant measurements. It is, therefore, crucial to have an approach that can ease up or even automate this process. In this report, we will introduce our solution.

1.2 Original Challenge Questions

Three challenge questions were posed by the challenge owner and advisors. They are:

- How to automatically identify peer-reviewed papers that contain data on observations of plastic in the ocean and on beaches?
- How to automatically parse that data into a database?
- How to geo-tag the plastic observations to location and time of sampling?

2 Approach

2.1 An Integrated Approach

To tackle the three challenge questions, we propose a integrated pipeline which begins with literature search and ends with a database of measurements (see Figure 1).

The first step of the pipeline is to gather records (e.g. journal articles, scientific reports, conference proceedings) that potentially contain measurements of plastics in the ocean and on beaches. The number of such records can easily go up to tens of thousands and obtaining the

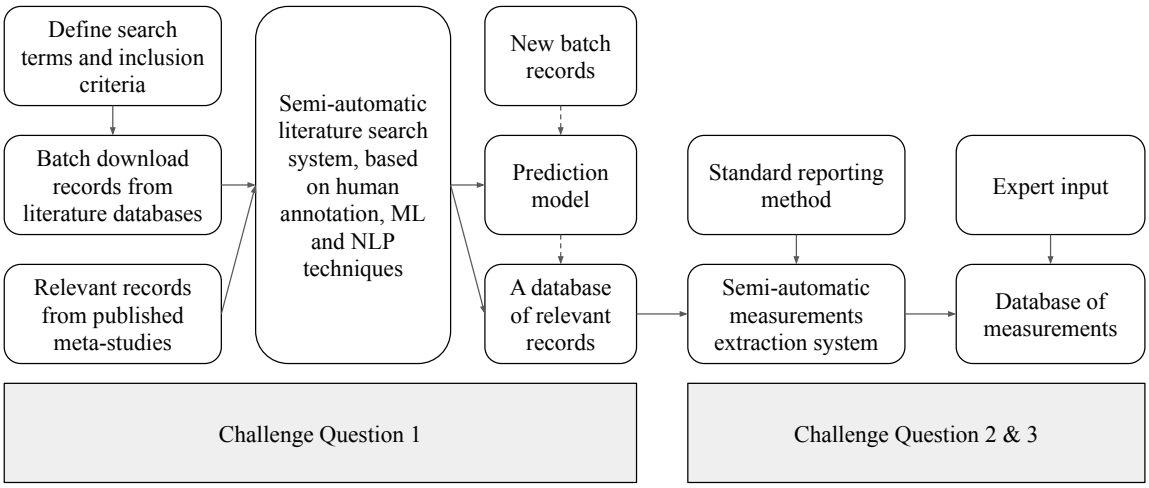


Figure 1: Proposed integrated pipeline for the challenge questions.

full texts of such a high number of records in a short amount of time is extremely difficult (if not impossible or illegal). Given this consideration, instead, we require only the meta information (e.g. title, abstract, keywords, publication venue, web links to full texts) of potentially relevant records in this step, which can be acquired much more readily. We will refer to them as meta records from now on.

Meta information of potentially relevant records can most conveniently come from **two sources**, the **first** of which are large literature databases, such as Scopus, Web of Science and PubMed. With a properly defined search string that corresponds to the research question of interest, databases will return many (but not exhaustively) potentially relevant records. Most databases allow (limited) batch downloading of meta-information of up to a few thousand returned records at a time. For instance, in Scopus, one can download abstracts of up to 2000 records every time. The **second** source are existing published meta-studies (e.g. meta analyses, systematic reviews) on the same research topic. They normally contain a list of included records, for which you can easily gather more meta information about. An additional advantage of using such records from existing meta-studies is that we do not have to annotate them again; they are readily available as data points for the **second step** of the pipeline: a semi-automatic literature search system based on human expertise and machine learning (ML) techniques.

This second step is the major contribution of our work, which will be detailed in the following section. The basic underlying idea is to train a supervised ML model that is capable of automatically predicting whether a record is relevant (i.e. containing plastic measurements in the ocean or on beaches) or not, based on limited meta-information available (i.e. titles, abstracts, keywords and publication venue). Naturally, any supervised ML model requires labelled data. Therefore, in this step, human annotation is needed (hence "semi-automatic" in the name), but several special tools are provided to speed up this process.

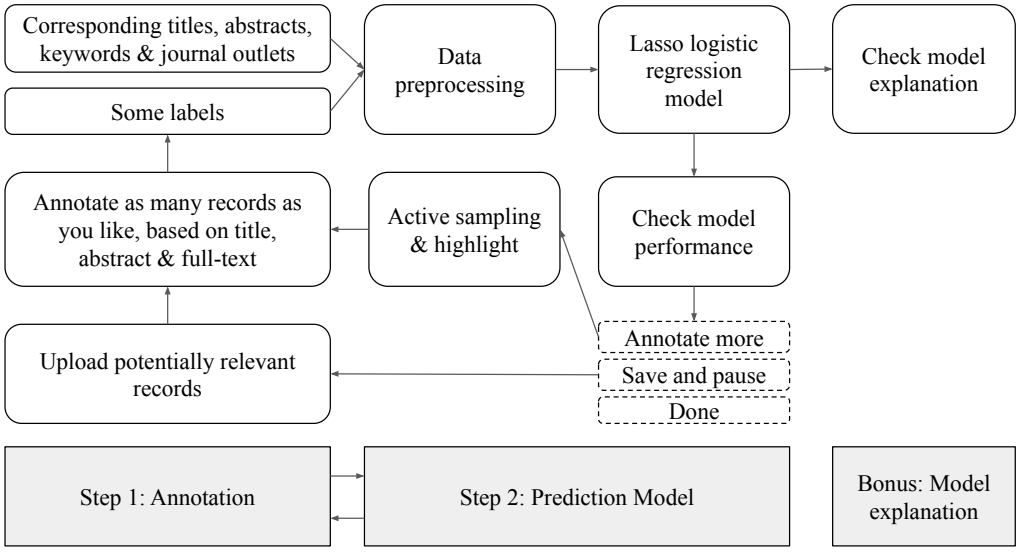
Ideally, the second step would produce **two outcomes**. The **first** one is a prediction model with satisfactory performance (as defined by the research question and the researcher). With this prediction model, we can quickly classify whether the remaining unlabelled records are relevant or not, resulting in a small, hopefully accurate selection of records that are very likely to be relevant. Consequently, the number of records we need to go through is much lower than if we had to manually go through all potentially relevant records. When the final screening is complete, we arrive at the **second** outcome of this step, a database of relevant papers. Furthermore, it is easy to keep the database up to date. For example, if one would like to update the database after one month, they can batch download potentially relevant meta records from literature databases that newly appeared in the last one month. Then, they can

give these meta records to the prediction model and screen the records predicted as relevant.

The two steps above are directly aimed at the first challenge question. The resulting database of relevant papers can then be used for tackling challenge question 2 and 3, which concern themselves with automatic extraction of plastic measurement data. We propose to use a semi-automatic system to extract measurement data from the full texts (i.e. PDF) of the relevant records, which results in a database of measurements. This system is considered semi-automatic because of two reasons. **First**, the amount of labelled data available for this system to learn is limited and therefore, human annotation will be needed. **Second**, the system is unlikely able to rely solely on ML-based prediction models, because of not only limited labelled data availability but also the fact that data extraction a hard problem (Jonnalagadda, Goyal, & Huffman, 2015). We expect that the system will make use of a combination of rule-based methods (such as using lexical patterns to search measurement data) and ML models. Furthermore, because of the heterogeneous ways of reporting data in publications, which makes (automatic) data extraction much more difficult, we recommend coming up with standard guidelines for reporting such data. Lastly, if the database of measurements has an online interface (such as a web application) and the interface allows users to submit data, then individual experts can contribute their own data to the database, subject to data validation by the management team.

To sum up, we believe that this proposed pipeline can effectively tackle the three challenge questions (i.e. automating the collection of measurements of plastics in the ocean and on beaches as much as possible). However, kindly note that due to limited time, we only focused on the first challenge question and our proposals regarding challenge question 2 and 3 remain conceptual.

2.2 Specific Approach for Challenge Question 1



Notes: The three dashed boxes after "Check model performance" represent three parallel choices.

Figure 2: Proposed semi-automatic literature search system (challenge question 1).

As mentioned previously, our main contribution to this challenge is a semi-automatic literature search system, which we detail in this section.

In this system, the first step is to upload data files containing potentially relevant meta records. Then comes data annotation, during which the researcher is presented with the meta information (like titles and abstracts) of a record at a time. The researcher needs to label this

record as relevant or irrelevant. If the researcher is uncertain, then he/she can use the full-text for better judgment.

Once enough papers have been annotated (the exact number is not important as you can always come back to annotate more), the annotation and the corresponding meta information like titles and abstracts are pre-processed and entered into a ML prediction model in Step 2. It is then important to check the performance of the prediction model trained on the annotated records. If you are not satisfied with the model and think that more data will improve the model, you can go back to Step 1 to annotate even more records. Note that now two additional features will become available to help you speed up your annotation even more. The first is "**active sampling**", which reorders the remaining, unannotated records in a way that those records that the prediction model is most uncertain about will be presented to you first. This has the advantage that the model will learn more efficiently and hence, you will likely need to annotate fewer records. The second feature is called "**highlight**". If you switch it on, for words that are considered by the prediction model to be crucial in predicting whether a record is relevant or not, they would be highlighted in the title and the abstract of every record you need to screen. This may help you identify key sentences more quickly. Now, you can iterate Step 1 and 2 until you have a satisfactory prediction model.

Alternatively, you are not yet satisfied with the model performance but you would like take a break and save your annotation results into your data file so far. You can do this. Whenever you are ready to go again, simply upload your updated data file and take it from there.

Additionally, you can check out how the model's predictions can be explained (e.g. what features are considered by the model to be important) and if that makes sense.

We have implemented this system as a Dash-based Python web application. See Section 3 for more information.

3 Results

3.1 Dash-based Python Web Application

To realise our proposed solution to challenge question 1 (i.e. automatically identifying papers that contain measurement observations of plastics in the ocean or on beaches), we developed a web application based on Dash. Dash is a Python framework for building web analytic applications, written on top of Flask, Plotly.js, and React.js. It is also open-source and easy to deploy. We used Dash of version 1.19.0.

The Appendix includes screenshots of the application that illustrate how the application can be used. Due to page limits, they are not shown here in the main paper body.

In Figure 3, you can see the initial interface of the application. You can see that under the main title "ML-assisted LITTERature Search", there is a horizontal bar containing three taps: "Step 1: Annotation", "Step 2: Prediction Model" and "Bonus: Model Explanation". These three taps correspond to the three components outlined in Figure 2. By clicking on a specific tab, you will be presented with a tab-specific interface. In the first two taps, each main interface is divided into three columns: left, middle and right.

After uploading your data file containing meta records, you are presented on the left column with a summary of your uploaded data, including the number of all records, that of papers already annotated as relevant, that of papers annotated as irrelevant and finally, that of unannotated records. In the middle column, you now see the title and abstract of a record that you still need to annotate. If you are not sure if the paper is relevant based on only the title and abstract, you can click on the "Full-text" badge, which will direct you on a new page to the site of the paper. This should provide you with enough information to make your annotation decision. As a side note, if you have an extension like "Library Access" installed on your browser (e.g. Chrome), you can visit the full text of the record more conveniently (as it automatically

signs you in via your institution).

On the right column, you can easily annotate any record by choosing one of "relevant", "irrelevant" and "uncertain" and then clicking the "Submit" button. Below the button you see the number of papers you have annotated.

Once you have some number of papers annotated, you can proceed by going to the next tap "Step 2: Prediction Model". You would then be presented with the interface shown in Figure 5. If you click the long blue button on the left, the prediction model will start running and an alert would pop up below the button to let you know whether the model is or has finished running. Once the model has finished running (see Figure 6), the middle column would be filled with the prediction performance scores of various evaluation metrics.

Now, if you are not happy about the model's prediction performance and you would like to improve it by annotating more data, then click on "I would like to annotate more" on the right column. This would present you with a light green alert below the three buttons, that you can now make use of two features to assist your annotation. That means, if you go back to the first tap "Step 1: Annotation", you would see two new radio buttons available on the right column (see Figure 7). They are "Active sampling" and "Highlight". If you activate "Active sampling", you will see that you are presented with a different record to annotate (see Figure 8), which is a record that the prediction model is most uncertain about. Actually, from now on, the system will present you with records in descending order of uncertainty. If you activate the "Highlight" feature, all the words considered by the prediction model to be important (for making the right prediction) would become highlighted (i.e. bolded) in the title and abstract.

Lastly, if you are curious about the importance of different words (i.e. features) used by the prediction model, you can go to the third tap "Bonus: Model Explanation" to check it out (see Figure 10).

3.2 Implementation Details

3.2.1 Meta records from Scopus and existing studies

One of the first issues to deal with in this challenge is to find a data set that contains potentially relevant meta records. For this, we discussed with the challenge owner and advisors and agreed on using Scopus for this purpose, with the following search string "(marine OR ocean OR sea) AND (plastic OR microplastic OR micro-plastic OR macroplastic OR macro-plastic OR nanoplastic OR nano-plastic OR mesoplastic OR meso-plastic)", for the period between 2008 and 2021 and limited to only journal articles and conference papers in English. This resulted in 7641 meta records.

For the development and testing of the Dash application, we used only the 2020 and 2021 records, totalling 1953. We, three researchers, then manually annotated 323 of them, cross-checked each other's annotation and eventually found 58 to be relevant.

In addition, we found 78 relevant records from an existing meta review publication (Kaandorp et al., 2020) and the advisory team's personal collections.

Combining all the records from above, we arrived at a development data set of 1991 meta records, with 95 of them annotated as relevant, 247 as irrelevant and 1649 still waiting to be annotated. Note that the previous numbers do not exactly add up to these final numbers because of overlapped records between different sources.

3.2.2 Data upload

To upload data to the system, two requirements need to be met. The first concerns the format of the file, which has to be either csv or Excel. The second requirement is that at least the following six columns (with exactly these names) should be included: Title, Keywords, Abstract, Venue, Link and Relevance.

3.2.3 Presenting a record

The title, abstract and quick access to full-text of a record is presented to the user.

3.2.4 Data preprocessing

For feature extraction, these four columns (Title, Keywords, Abstract and Venue) are concatenated into one single feature column. That is to say, each record (i.e. row) has now a long string that contains information about its title, keywords, abstract and publication venue (if there is no missing data). Then, we tokenized the strings, lower-cased the characters, removed common English stop words and removed words that have either a very high document frequency or a very low one. Next, we randomly split all the annotated records into a training set (80%) and a test set (20%). Within the training set, we built a vocabulary corpus and took from it the 100 most common words as new features. The values of these 100 features for a given record are the corresponding word counts. Using these 100 features, we built a feature matrix of token counts for the training and the testing data, respectively.

3.2.5 Imbalanced data

When retrieving candidate articles from a database, researchers will want to use a query that is broad enough not to leave potentially relevant articles out. As a result, the fraction of relevant articles in the corpus retrieved is typically rather small. For example, when considering four systematic reviews of a diverse range of fields, researchers noted that only between 0.66% and 4.84% of the articles screened ended labelled as relevant (van de Schoot et al., 2021). For our manually labelled development data set about 27.7% of articles were marked as relevant. Still, this is likely an overestimation of the actual proportion on all retrieved records from the database because of our inclusion of additional relevant articles from existing meta studies and other sources.

Training models to predict an outcome variable with such an unequal prevalence on its classes is a well known methodological challenge in the field of machine learning. For example, a model trained classify articles as relevant or irrelevant with a data containing only 0.66% of relevant articles would achieve an accuracy of 99.34% just by labelling all the cases as irrelevant.

To avoid a classifier that achieves a high accuracy while failing to identify the relevant papers, there are various strategies that can be followed. For example, several studies have showed the potential of resampling strategies to deal with this issue (O'Mara-Eves, Thomas, McNaught, Miwa, & Ananiadou, 2015). They achieved balancedness on the training set by either oversampling the minority class, undersampling the majority class, or a combination of the two (Krawczyk, 2016). Yet, resampling methods can be quite computationally demanding for large and highly imbalanced data (Krawczyk, 2016). Here, we follow a different approach based on the inclusion of sample weights to the training phase of the model (Krawczyk, 2016). These sample weights are straightforward to calculate and include in the model, as they consist in the inverse of the probability of inclusion of the class (e.g. for a data with a class imbalance of 1:10, each case of the minority class receives 10 times more weight than each case of the majority class). Consequently, the model will penalize more a misclassification of the minority class than one of the majority class (it will have a larger weight).

3.2.6 Prediction models and performance

We used Lasso logistic regression for the final implementation of the prediction model in the application. Compared to regular logistic regression, it has the advantage of being able to deal with high-dimensional data (especially when our sample size is not much larger than 100, namely, the number of features), showing lower model variance (hence better prediction) and

often resulting in a simpler, more interpretable model. It is also more light-weight, generally faster than more complex models like random forests and XGBoost.

In Table 1, you see that the prediction performance of Lasso logistic regression on the test data set is comparable (if not better than) the other models. Noticeably, the Lasso logistic regression model also has a very balanced performance across all metrics.

Table 1: Model evaluation metrics.

Model	F1	AUC	Precision	Recall
llr	0.848	0.927	0.824	0.875
gbm	0.813	0.960	0.813	0.813
xgboost	0.769	0.820	0.750	0.789
ranger	0.640	0.940	0.889	0.500

Note. Models trained with sample weights; metrics calculated on test set; models: llr (lasso logistic regression), gbm (gradient boosting), xgboost (extreme gradient boosting), & ranger (random forest).

3.2.7 Active sampling

As aforementioned, the application introduces an active sampling feature to help users annotate records. The idea is to rank the unannotated records by how uncertain the prediction model is about them and present the least certain records to the users first. In this way, the model can learn more efficiently.

This idea actually comes from the field of active learning and this sampling strategy is called uncertainty sampling. Combining these two terms gives us "active sampling".

3.2.8 Model explanation and the highlight feature

To help the users understand how the model makes prediction, we used so-called permutation feature importance scores (Altmann, Toloşi, Sander, & Lengauer, 2010). The basic idea is that by randomly permuting the values of a feature, if the prediction performance of the model goes down, then the feature has to be important. Otherwise, the feature is not important. The feature importance score is then calculated as the percentage increase in prediction error. The higher the score is, the more important a feature is. A score equal to or below zero indicates that the feature is not important at all.

In the "Bonus: Model Explanation" tab of the application, we show a figure with the most important features (i.e. words), defined as having a importance score > 0.01. The very same "important" words are used by the "Highlight" feature.

4 Summary and conclusions (max 2 pages)

Researchers make constant attempts to measure plastic concentrations in the ocean and on beaches and publish their findings in scientific venues. Gathering all such relevant studies and combining their findings can help the scientific community (and the society at large) to stay informed about the status-quo of global plastic pollution and to estimate the future movement of plastic litter. However, these studies, very much like their study subject (i.e. plastic litter), are scattered in the ocean of scientific publications. The measurement data they contain are also often hard to locate and extract.

Hence, this data challenge aims at automating 1) the collection and identification of relevant studies that contain measurements of plastic litter in the ocean and on beaches; 2) the extraction of the measurements and their meta data.

4.1 Main advancements

We have made two main contributions. **First**, we proposed a conceptual, integrated pipeline to achieve these two goals (as shown in Figure 1), which leverages both human expertise and machine intelligence. **Second**, in particular, we developed a web application prototype that is a realisation of the first part of this integrated pipeline. Specifically, this web application makes use of ML and active learning to automate the collection and identification of relevant studies as much as possible. We showed that this approach achieved reasonable and explainable performance on the development data set.

It is true that our web application is not the first of its kind. There are existing tools such as ASReview (van de Schoot et al., 2021) and Rayyan (Ouzzani, Hammady, Fedorowicz, & Elmagarmid, 2016), both of which use similar ML techniques to speed up the search of relevant literature. However, our application still has four unique advantages compared to other tools. **First**, it is a customised solution to the current, specific challenge at hand. It can be easily extended in the future to also include features outlined in the second half of the proposed pipeline (see Figure 1), such as a data extraction system and a database of measurements. **Second**, our tool provides direct access to the full text of a record, which no other tool (to our knowledge) can deliver. This makes it very convenient for when researchers need more information than the title and the abstract to decide whether a paper is relevant. **Third**, we provide model explanation to enable our users to understand their prediction model better. **Lastly**, we implemented the very "Highlight" feature to help users speed up annotation, which, to our knowledge, no other tools currently provide.

4.2 Limitations and further steps

The main limitation of our work is that while focusing on the first challenge question, we leave the solution to the second and third challenge question on a mostly conceptual level. Ideally, we would have liked to also develop a working prototype for the data extraction part of the challenge. For the web application we developed, we would like to subject it to more testing before it can be officially deployed. We would also like to conduct some user tests, to investigate to what extent the features we implemented (e.g. access to full texts, active sampling, highlight, prediction models) contribute to the speed and accuracy of annotation.

4.3 Role of complex systems science

Despite this being a complexity science data challenge, we did not make explicit use of any complexity science knowledge or tools. However, we do see the connection between this data challenge and complexity science. For instance, the main subjects of our study are written documents, which can arguably be seen as complex systems, as they exhibit some typical features of complex systems. For instance, documents are composed of many components (i.e. characters, words, symbols) that interact with each other (particularly during the writing process) in a non-linear way. Different documents, via their authorship and citations, can form networks. In future work, we can consider making use of, for instance, citation networks and the embedded knowledge in the texts to further help with the automatic identification of relevant papers and data extraction.

Acknowledgements

We very much appreciate the engagement and valuable input of Dr. Erik van Sebille, Mikael Kaandorp, Darshika Manral and Sophie Schmiz during this challenge.

Data & code availability

All the code and data have been made available at *github.com/fqixiang/LITTERatureSearch*. The Dash-based web application can be accessed at *dash-complexaton-5.herokuapp.com*. Kindly note that the application has not gone through rigorous testing and therefore, bugs may occur. We recommend trying out the application by using the same data set we used, which can be downloaded at
github.com/fqixiang/LITTERatureSearch/blob/main/Qixiang/data/data_complete_test.csv.

References

Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010, May). Permutation importance: a corrected feature importance measure. *Bioinformatics*, *26*(10), 1340–1347. Retrieved 2021-04-06, from https://doi.org/10.1093/bioinformatics/btq134 doi: 10.1093/bioinformatics/btq134

Barnes, D. K. A., Galgani, F., Thompson, R. C., & Barlaz, M. (2009, July). Accumulation and fragmentation of plastic debris in global environments. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1526), 1985–1998. Retrieved 2021-04-06, from https://royalsocietypublishing.org/doi/abs/10.1098/rstb.2008.0205 (Publisher: Royal Society) doi: 10.1098/rstb.2008.0205

Booth, A. M., Kubowicz, S., Beegle-Krause, C., Skancke, J., Nordam, T., Landsem, E., & Jahren, S. (2017). Microplastic in global and Norwegian marine environments: Distributions, degradation mechanisms and transport. , 149.

Chiba, S., Saito, H., Fletcher, R., Yogi, T., Kayo, M., Miyagi, S., ... Fujikura, K. (2018, October). Human footprint in the abyss: 30 year records of deep-sea plastic debris. *Marine Policy*, *96*, 204–212. Retrieved 2021-04-06, from https://www.sciencedirect.com/science/article/pii/S0308597X17305195 doi: 10.1016/j.marpol.2018.03.022

Cózar, A., Echevarría, F., González-Gordillo, J. I., Irigoien, X., Úbeda, B., Hernández-León, S., ... Duarte, C. M. (2014, July). Plastic debris in the open ocean. *Proceedings of the National Academy of Sciences*, *111*(28), 10239–10244. Retrieved 2021-04-06, from https://www.pnas.org/content/111/28/10239 (Publisher: National Academy of Sciences Section: Biological Sciences) doi: 10.1073/pnas.1314705111

Goldstein, M. C., Titmus, A. J., & Ford, M. (2013, November). Scales of Spatial Heterogeneity of Plastic Marine Debris in the Northeast Pacific Ocean. *PLOS ONE*, *8*(11), e80020. Retrieved 2021-04-06, from https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0080020 (Publisher: Public Library of Science) doi: 10.1371/journal.pone.0080020

Jonnalagadda, S. R., Goyal, P., & Huffman, M. D. (2015, June). Automating data extraction in systematic reviews: a systematic review. *Systematic Reviews*, *4*(1), 78. Retrieved 2021-04-06, from https://doi.org/10.1186/s13643-015-0066-7 doi: 10.1186/s13643-015-0066-7

Kaandorp, M. L. A., Dijkstra, H. A., & van Sebille, E. (2020, October). Closing the Mediterranean Marine Floating Plastic Mass Budget: Inverse Modeling of Sources and Sinks. *Environmental Science & Technology*, *54*(19), 11980–11989. Retrieved 2021-04-06, from https://doi.org/10.1021/acs.est.0c01984 (Publisher: American Chemical Society) doi: 10.1021/acs.est.0c01984

Krawczyk, B. (2016, November). Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, *5*(4), 221–232. Retrieved 2021-04-06, from https://doi.org/10.1007/s13748-016-0094-0 doi: 10.1007/s13748-016-0094-0

Law, K. L., Morét-Ferguson, S. E., Goodwin, D. S., Zettler, E. R., DeForce, E., Kukulka, T., & Proskurowski, G. (2014, May). Distribution of Surface Plastic Debris in the Eastern Pacific Ocean from an 11-Year Data Set. *Environmental Science & Technology*, *48*(9), 4732–4738. Retrieved 2021-04-06, from https://doi.org/10.1021/es4053076 (Publisher: American Chemical Society) doi: 10.1021/es4053076

Morét-Ferguson, S., Law, K. L., Proskurowski, G., Murphy, E. K., Peacock, E. E., & Reddy, C. M. (2010, October). The size, mass, and composition of plastic debris in the western North Atlantic Ocean. *Marine Pollution Bulletin*, *60*(10), 1873–1878. Retrieved 2021-04-06, from https://www.sciencedirect.com/science/article/pii/S0025326X10003267 doi: 10.1016/j.marpolbul.2010.07.020

Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016, December). Rayyan—a web and mobile app for systematic reviews. *Systematic Reviews*, *5*(1), 210. Retrieved 2021-04-06, from https://doi.org/10.1186/s13643-016-0384-4 doi: 10.1186/s13643-016-0384-4

O'Mara-Eves, A., Thomas, J., McNaught, J., Miwa, M., & Ananiadou, S. (2015, January). Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Systematic Reviews*, *4*(1), 5. Retrieved 2021-04-06, from https://doi.org/10.1186/2046-4053-4-5 doi: 10.1186/2046-4053-4-5

Schlining, K., von Thun, S., Kuhnz, L., Schlining, B., Lundsten, L., Jacobsen Stout, N., ... Connor, J. (2013, September). Debris in the deep: Using a 22-year video annotation database to survey marine litter in Monterey Canyon, central California, USA. *Deep Sea Research Part I: Oceanographic Research Papers*, *79*, 96–105. Retrieved 2021-04-06, from https://www.sciencedirect.com/science/article/pii/S0967063713001039 doi: 10.1016/j.dsr.2013.05.006

Sebille, E. v., Wilcox, C., Lebreton, L., Maximenko, N., Hardesty, B. D., Franeker, J. A. v., ... Law, K. L. (2015, December). A global inventory of small floating plastic debris. *Environmental Research Letters*, *10*(12), 124006. Retrieved 2021-04-06, from https://iopscience.iop.org/article/10.1088/1748-9326/10/12/124006/meta (Publisher: IOP Publishing) doi: 10.1088/1748-9326/10/12/124006

van de Schoot, R., de Bruin, J., Schram, R., Zahedi, P., de Boer, J., Weijdema, F., ... Oberski, D. L. (2021, February). An open source machine learning framework for efficient and transparent systematic reviews. *Nature Machine Intelligence*, *3*(2), 125–133. Retrieved 2021-04-06, from https://www.nature.com/articles/s42256-020-00287-7 (Number: 2 Publisher: Nature Publishing Group) doi: 10.1038/s42256-020-00287-7

Wayman, C., & Niemann, H. (2021). The fate of plastic in the ocean environment – a minireview. *Environmental Science: Processes & Impacts*, *23*(2), 198–212. Retrieved 2021-04-06, from https://pubs.rsc.org/en/content/articlelanding/2021/em/d0em00446d (Publisher: Royal Society of Chemistry) doi: 10.1039/D0EM00446D

Appendix

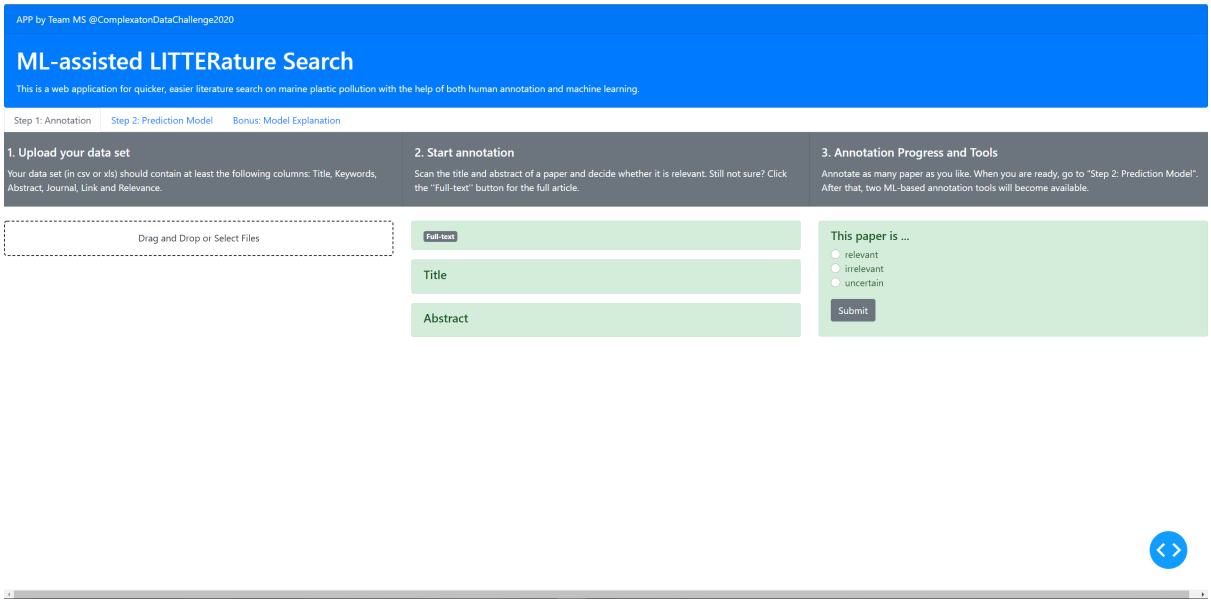


Figure 3: The initial interface of the Dash-based web application.

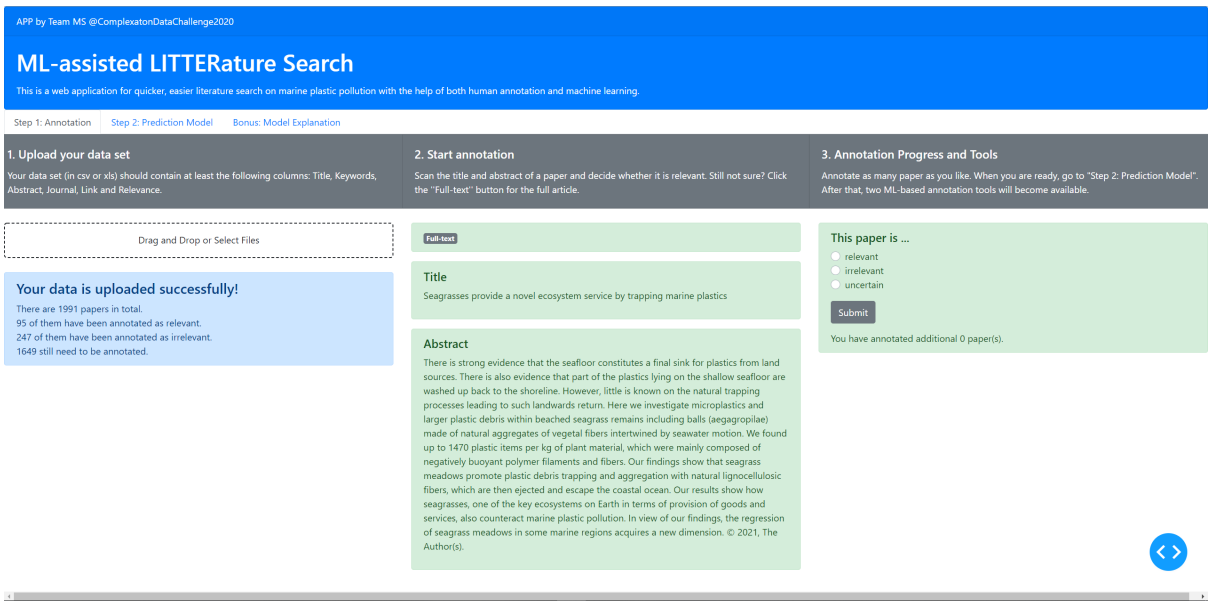


Figure 4: The interface after data upload (still in Tab 1).

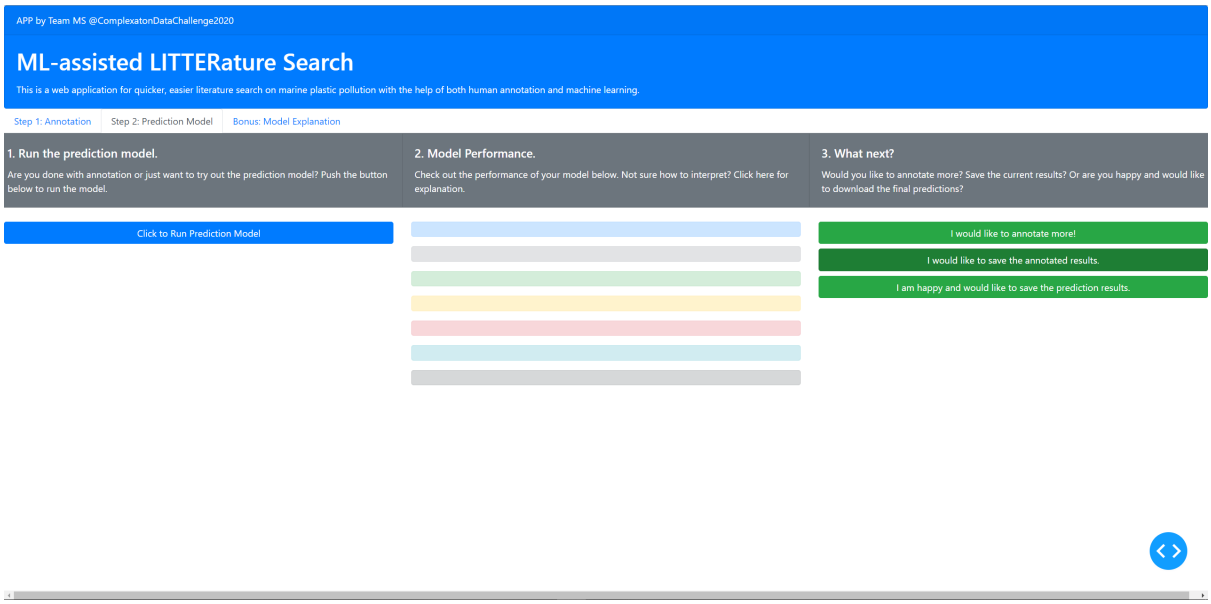


Figure 5: The initial interface of Tap 2, before running a model.

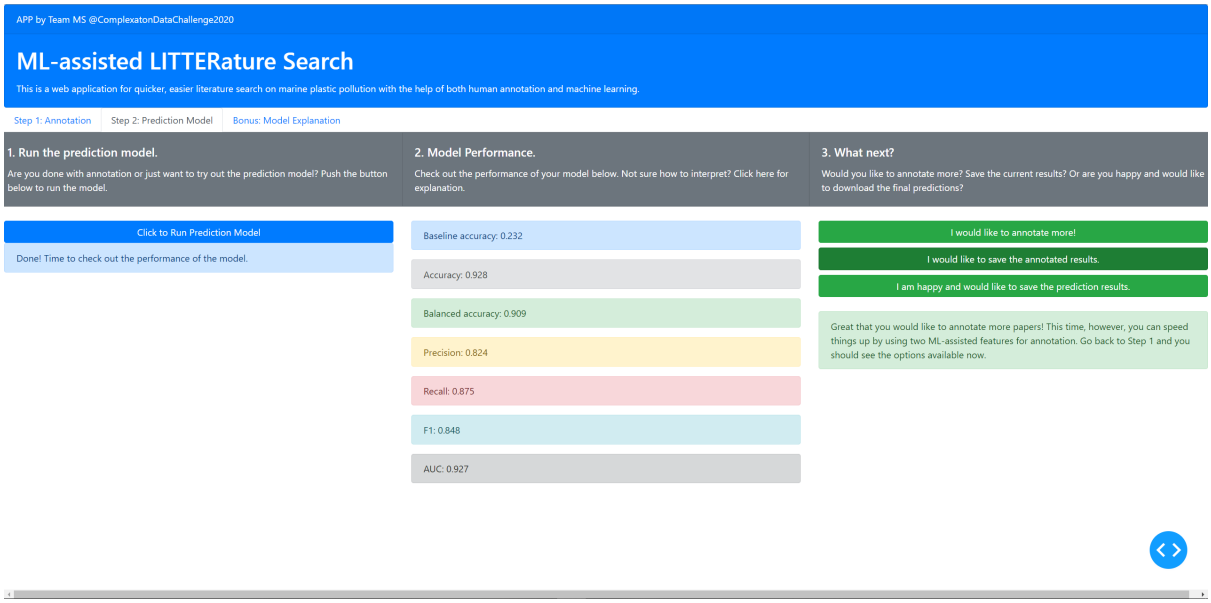


Figure 6: The interface of Tap 2, after running a model and clicking "I would like to annotate more!".

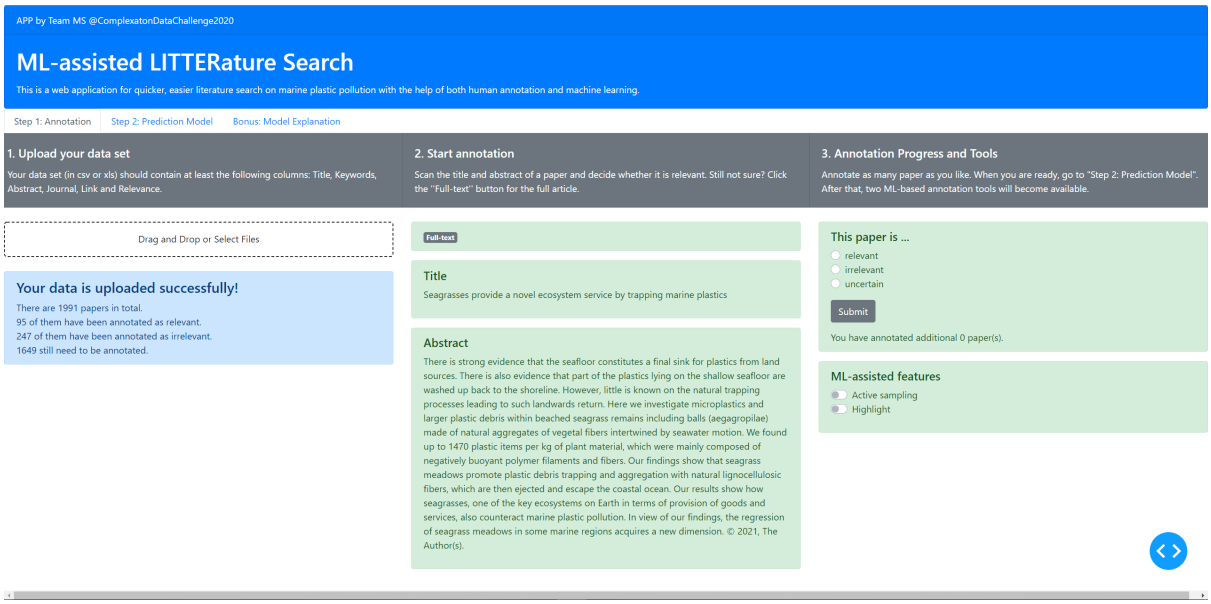


Figure 7: The interface of Tap 1, after running a model and clicking "I would like to annotate more!".

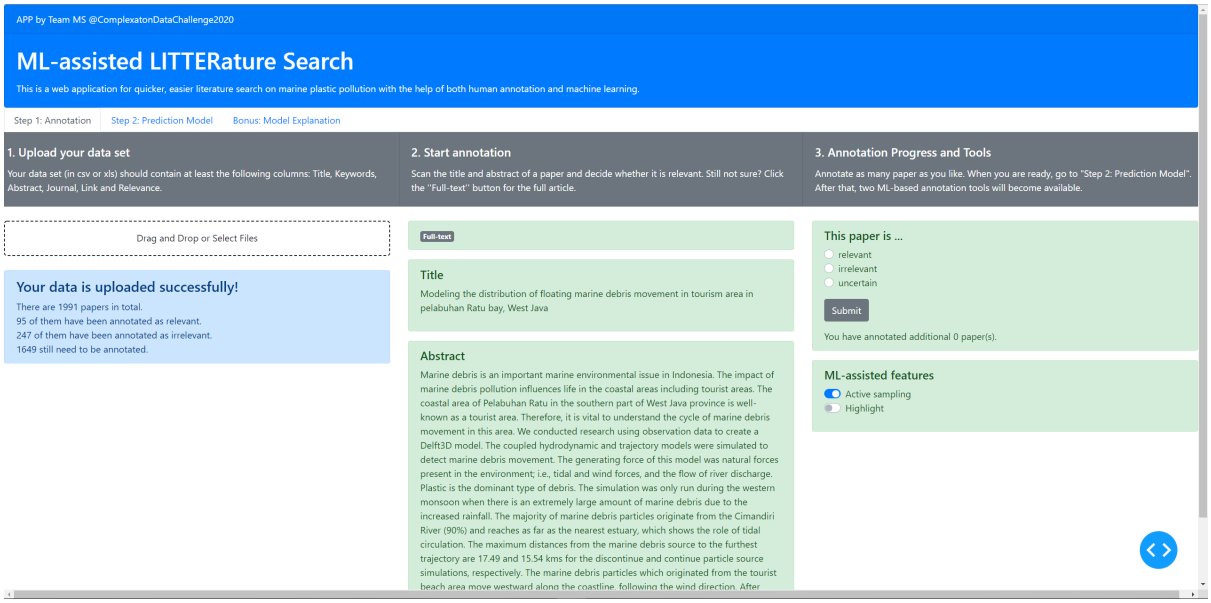


Figure 8: The interface of Tap 1 with active sampling feature

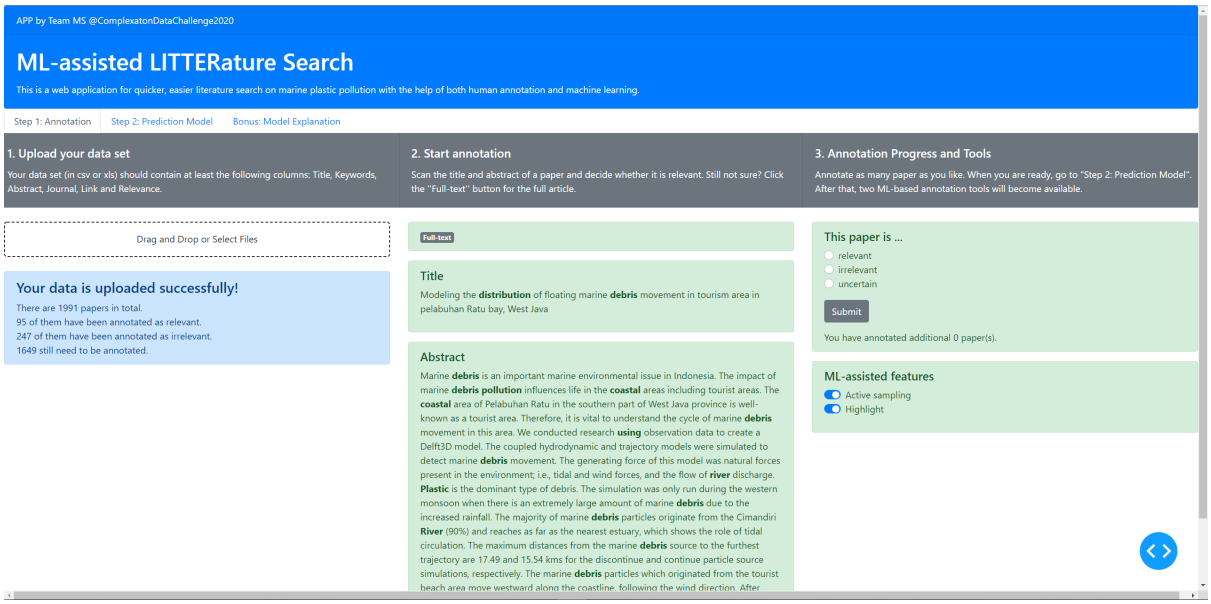


Figure 9: The interface of Tap 1 with both active sampling and highlight feature

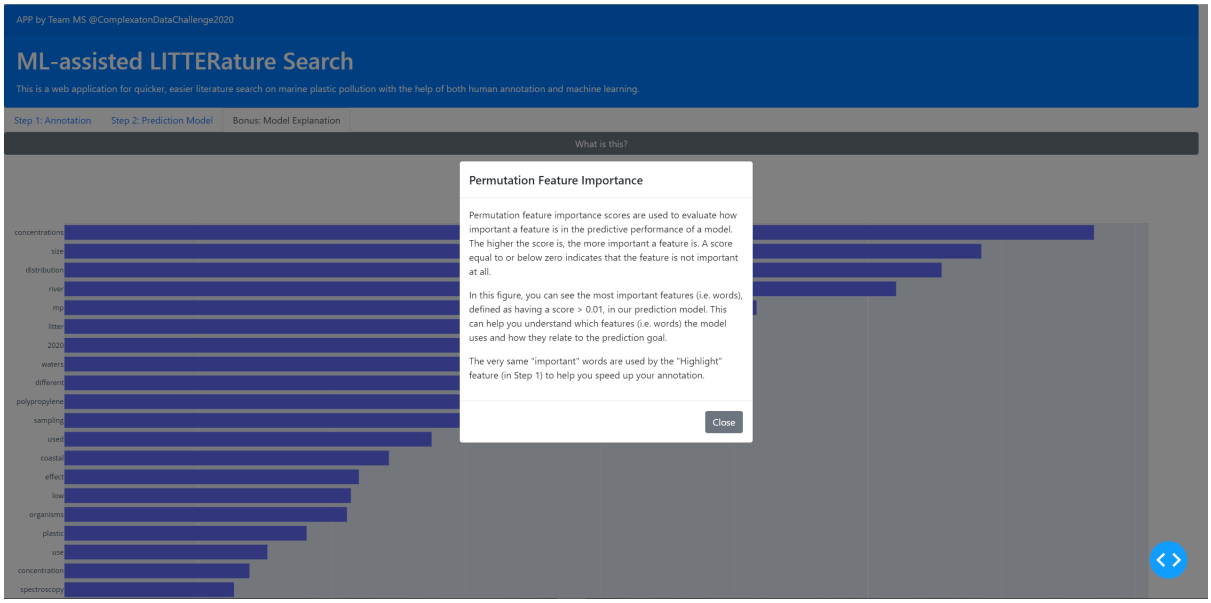


Figure 10: The interface of Tap 3 "Bonus: Model Explanation" with permutation feature importance