Capstone Project | Machine Learning Engineer Nanodegree
Project Proposal
Jorge Saldivar

Automatically Identifying Evidence of Social Impact in Research Documents

Background

From some time now, there has been an increasing interest in society to understand the scientific and social impact of the financial resources invested in research. In this context, initiatives like Open Science, Open Access, and Open Data have emerged to satisfy this demand. As a result, and today as never before, the public has access to scientific knowledge disseminated through new and traditional media channels and available in online and offline spaces of socialization.

Scientometrics scholars have been studying for several years now different alternatives for measuring the scientific impact of research, and there is already an ample consensus on well-established methods to assess the scientific contribution of investigations. However, there is still not an agreement on techniques to evaluate the societal impact of research, and several proposals have emerged in the last years. SIOR is one of these approaches proposed by Flecha [1] to measure the social impact of research. The method requires to process all of the documentation produced by the research project (technical reports, scientific articles, white papers, book chapters) to identify inside them evidence of social impact. Once the evidence has been found, Flecha proposes to evaluate it through the following five indicators:

1. Connection with the Sustainable Development Goals (SDG) of the United Nations, or EU2020 targets, or other similar officially defined social goals;
2. Percentage of improvement achieved in relation to the existing situation;
3. Replicability of the impact, i.e., whether the evidence reports impact in more than one social context;
4. Sustainability of the impact, i.e., whether the reported evidence show to be sustainable over time;
5. Source of evidence, i.e., whether the evidence has been reported in a scientific article or government official document.

After evaluating the identified evidence through the indicators, SIOR proposes a calculation to obtain a score from 1 to 10 that indicates the social impact of the research project. Please refer to the SIOR website for more information about the indicators and on how the methodology is applied.

Problem Statement

SIOR has been initially proposed to be a self-assessment methodology to measure the social impact of research. That is to say, researchers have to apply SIOR to understand the impact of their investigations. However, the approach would impose substantial human effort to official agencies that might be interested in measuring the social impact of their funded research projects. This limitation because it requires processing vast volumes of unstructured and textual information produced by the research to find evidence of impact before applying SIOR.

Solution

In this project, I propose to leverage machine learning and natural language processing techniques to build a text classifier that automatizes the processing and identification of evidence of social impact in research documents. The proposal aims to solve a classification problem in which the model takes a sentence as input and produces as output a binary answer (1=True, 0=False) that states whether the sentence contains evidence of social impact. By automatically identifying proof of impact in documents produced by research projects, the solution will help research funding agencies and governmental institutions in the application of SIOR methodology to assess the social impact of their funded projects.

Among all research fields, this project focuses on Medical, Health, and Biological science because the plan is to apply the resulting solution in the evaluation of the social impact of the research projects of the Spanish National Institute of Bioinformatics (INB by its Spanish Acronym). INB is an institution that conducts medical and biological investigations.

Datasets

Training a machine-learning algorithm to automatically identify evidence of social impact in text documents requires having a dataset with examples of sentences that provide evidence of impact. To the best of my knowledge, there are no datasets of this type, at least I could not find them. So, I have to build my datasets of sentences with evidence of impact. In doing so, I am going to leverage on the summaries of the societal impact of research published by the Research Excellence Framework (REF)—REF is the evaluation system used in the United Kingdom to measure the quality of research of their higher education institutions.

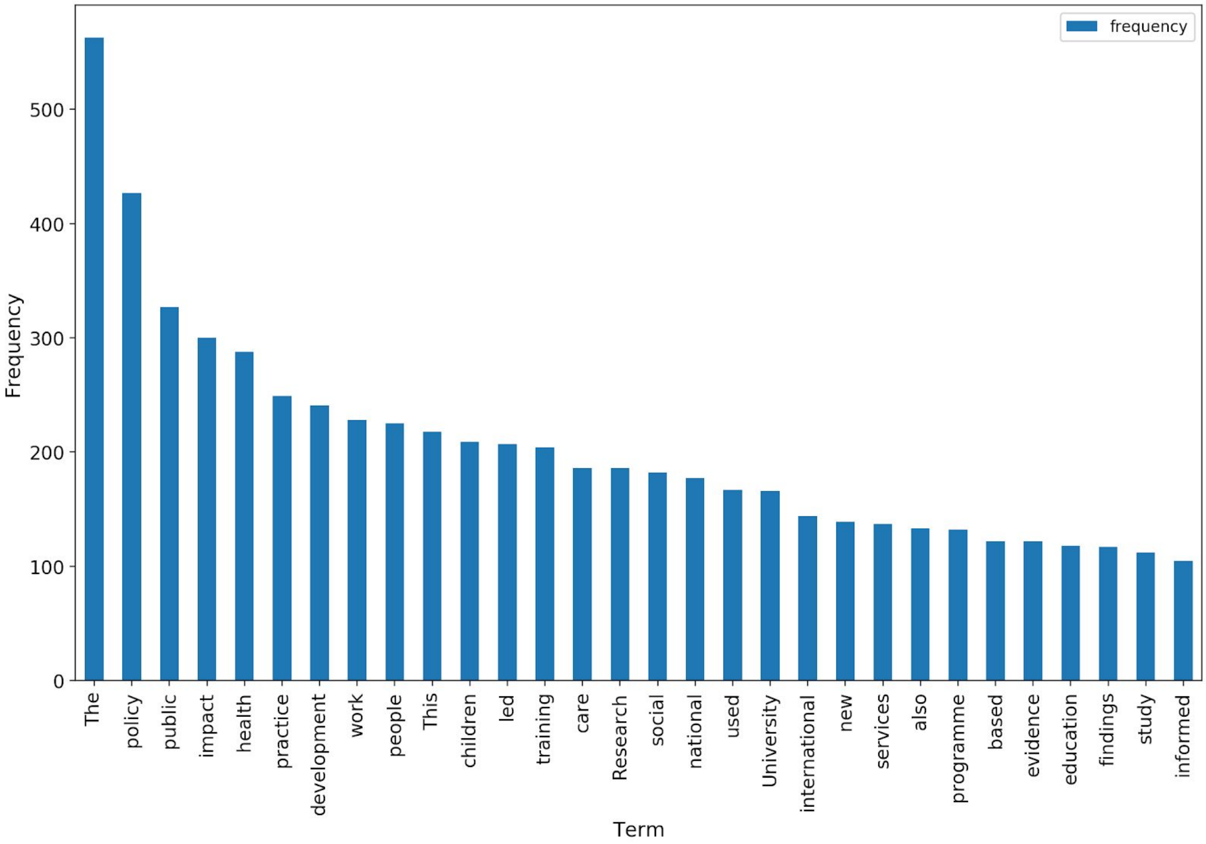
Summaries of the social impact of Medical, Health, and Biological research conducted in the United Kingdom have been downloaded from the REF website (e.g., summaries of medical and health science can be accessed here). The dataset containing 571 summaries of the social impact of biological, medical, and health science research are available at the repository of this project here. Summaries, which are a few paragraph descriptions of how the research has impacted the society, have been split into sentences, resulting in 2,347 sentences that have been assumed all of them contain evidence of social impact. Before splitting, I have removed the dot after abbreviations that are commonly employed in academic writing (e.g., i.e., w.r.t., a.o.) to avoid the sentence tokenizer to start a new sentence after these abbreviations.

These sentences with evidence of social impact have been complemented with general sentences that are commonly available in scientific documents in the field of Health and Biology. Here, I have used the dataset of 29,437 articles of health and biology published as part of the publication SparkText: Biomedical Text Mining on Big Data Framework to extract sentences that are usually part of health and biology academic articles. The dataset of almost 30,000 articles can be downloaded from here, each record in the dataset contains the full text of an article. The text of the 29,437 articles has been split into sentences, resulting in 2,318,588 sentences. Then, I have randomly selected 3,000 sentences to complement the sentences that contain evidence of social impact. The number 3,000 was chosen to have approximately 5,000 sentences in the combined dataset. More than 3,000 sentences can be selected, but this will increase the unbalance in the final dataset. The big assumption in this part of the process is that the selected "general" sentences do not contain evidence of social impact. Only grammatically complete sentences have been included in the selection. For this project, sentences are considered complete if they have at least two nouns (subject, object) and a verb. The process of building the dataset of general sentences has been documented in the notebook dataset preparation. The dataset of the 3,000 general sentences is available at the repository of the project.

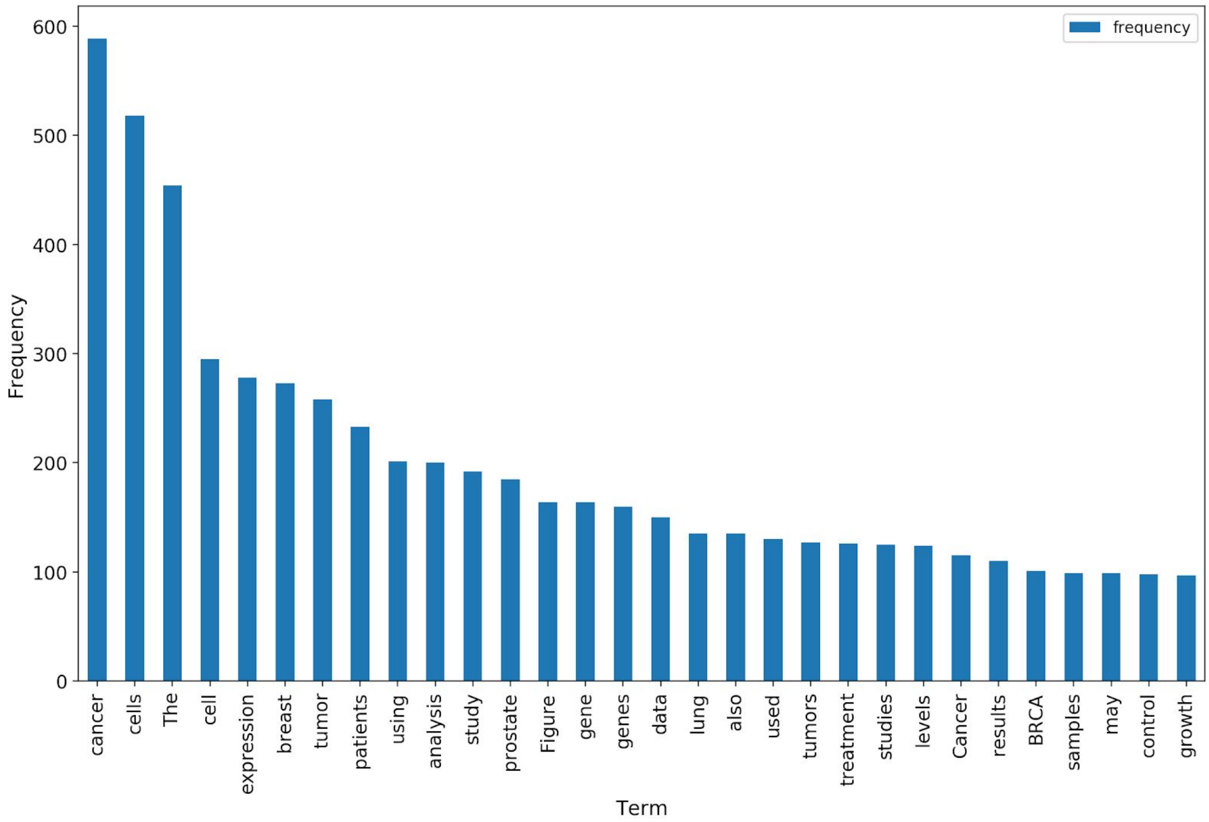
After applying some cleaning tasks (i.e., remove HTML tags, delete URLs, get rid of punctuations as well as extra white spaces) to both datasets (i.e., the dataset containing sentences with evidence of social impact and the dataset of general sentences), they have been merged into a single dataset.

Data Exploration

The combined dataset has been explored by, first, computing the most frequent terms in sentences that contain evidence of social impact and, then, by calculating the most frequent terms in general sentences. Figure 1 shows the 30 most frequent terms in both types of sentences.



(a) Most frequent terms in sentences with evidence of social impact



(b) Most frequent terms in general sentences

Figure 1. The 30 most frequent terms in sentences containing evidence of social impact and in general sentences.

Figure 1 shows that sentences with evidence of social impact (a) and general sentences (b) are expressed using very distinctive languages. There is little overlap between the most frequent terms in sentences with evidence of social impact (e.g., public policy, health, national, development) and the most frequent terms in the general sentences. Terms in the latter are more related to the medical and biological terminologies, such as cancer, cells, patients, tumor, or therapy. The low interception between the most frequent terms in the two types of sentences demonstrates the potential of the data to build a robust classifier that can clearly distinguish between general sentences and sentences containing evidence of social impact.

Features Generation

Features have been generated through two different strategies. On the one hand, features have been constructed by extracting lexical, syntactic, and semantical information of sentences. On the other hand, features have been built by converting sentences into vectors of tokens.

Regarding the first strategy, I have created the following five features to represent different lexical and syntactic characteristics of the sentences:

- Word count of sentences: total number of words in sentences;

- Character count of sentences: total number of characters in sentences;
- Average word density of sentences: average length of words in sentences;
- The upper case count in sentences: total number of upper count words in sentences;
- Title word count in sentences: total number of title case words in sentences.

Apart from the lexical and syntactic features, the sentiment of sentences has been computed. The reasoning behind this feature is that sentiments can be a promising feature to discriminate between the two types of sentences since, as it is shown in Figure 2, sentences with evidence of social impact have, on average, a more positive attitude than the general sentences. Vader sentiment analyzer has been used to calculate the sentiment of sentences.

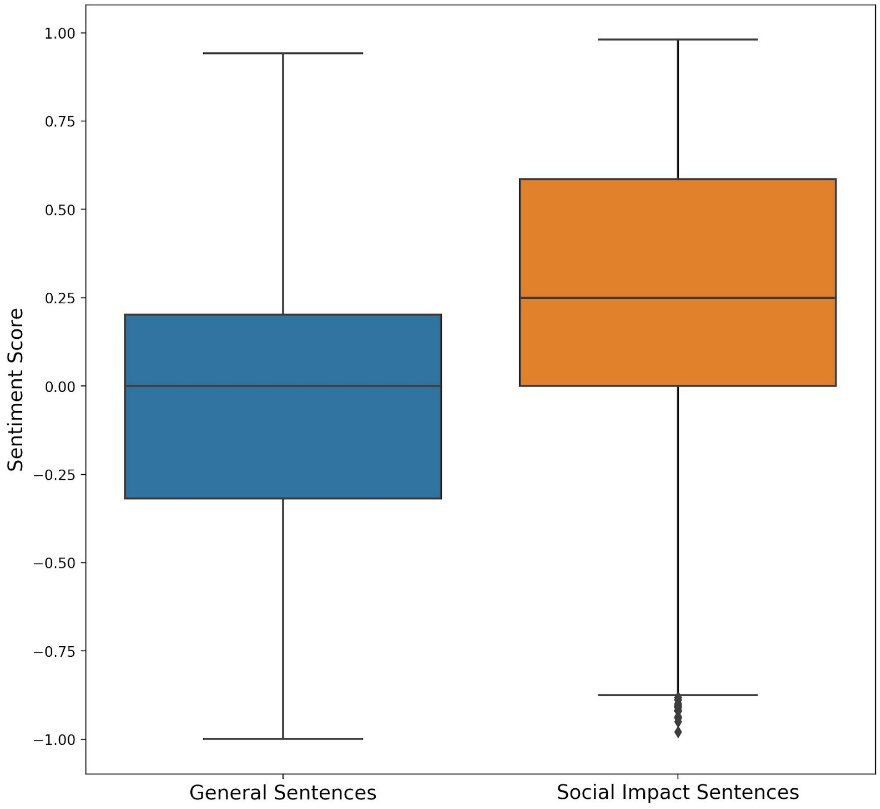


Figure 2. Distribution of sentiment scores of sentences with evidence of social impact and general sentences

After computing the lexical, syntactic, and semantic features, the dataset has been split into train and test. Twenty percent of the sentences have been held for testing while the rest have been reserved for training. Sentences in both the train and test sets have been vectorized, converting them into arrays of tokens. Before vectorizing, sentences have been preparing to apply the following combinations of text processing tasks:

1. Sentences are tokenized and stop words, as well as URLs, removed. Stemming is also used;
2. Sentences are tokenized, stop words and URLs are removed, and lemmatization is applied.

Sentences in the train and test sets have passed through the tasks described in (1) and (2), generating four sets of sentences, two pairs of train-test sets.

Later, the sets of sentences have passed through various transformations that include combinations of different bag-of-words representations. Term-frequency (TF) and term frequency-inverse document frequency (TF-IDF) have been employed to vectorize sentences with five different values for the maximum features parameter (i.e., 500, 800, 1000, 2000, and 5000) and four different ngram ranges, namely (1,1), (1,2), (1,3), and (2,3). In total, 40 different transformations (i.e., two vectorized representations (TF and TF-IDF) * five values of maximum features * four ngram ranges) have been applied to the sentences pre-processed through the two groups of tasks described above. The transformed sentences together with the syntactic and semantic features have been saved into 160 train and test binary files (80 files containing sentences of the train set and 80 files with sentences of the test set).

Performance Metric

I chose F1 as the performance metric to evaluate models. F1 combines both metrics Precision and Recall providing an adequate compromise between acceptable coverage and the correct identification of impact sentences.

Model Building

In the process of building the model, five algorithms have been trained. The decision on the algorithms to be used was based on empirical evidence that reports the algorithms that are known to perform well on unbalanced, relatively-small, and textual datasets [2]. In this sense, I have tried the following algorithms: Gaussian Naive-Bayes (NB), Support Vector Machine (SVM), Logistic Regression (LR), Random Forest (RF), and Gradient Boosting (GB). Each algorithm has been trained with their default hyper-parameters and using cross-validation on the 80 different train sets. This part of the process produces 400; this is, 80 models per each of the five algorithms. Then, the best performing model of each algorithm has been selected based on the mean of F1 scores obtained from the cross-validation process. Figure 3 depicts the mean F1 scores of each of the best performing models.

Later, I have conducted a hyper-parameters tuning of all of the best performing models, excluding the Naive-Bayes model. I have decided to leave unchanged the only parameter var_smoothing of the Gradient Naive-Nayes since I found the algorithm performs pretty well with the default parameter. The parameters of the rest of the models have been optimized using Random Grid Search with 100 interactions.

Following hyper-parametrization, learning curves have been plotted to analyze the overfitting effect on the train set and the effect of the train size on the performance metric. Figure 4 shows the learning curves of each model. In Figure 4 (a), we can see that scores in the learning curve of the NB model decrease as the size of the training set increases. This result indicates that the model is learning well on the train data. On the other hand, the validation curve increases as more training examples are available, and it seems that with more data curves will converge at some point. Learning curves of the SVM, which are outlined in Figure 4 (b), indicate that the scores remain constant as the size of the training set increases. This behavior is a clear sign of high-variance and suggests that the model overfits on the training data. Scores in the learning curve of the LR model decrease as the number of train examples increases see Figure 4 (c). This result shows that the model does not overfit on the train data (low variance). On the other hand, the validation curve increases as more training examples are available, which tells us that the model is correctly learning from the training examples. The small gap between curves is a sign of low variance; however, the large shadows around curves are an indication of significant variability in the scores. The learning curves of the models RF and GB, which are shown in Figures 4 (d) and 4 (e), unveil in both a clear sign of overfitting on the train data since training scores are at maximum regardless of the train size. The large gap between curves in both models is a strong sign of high-variance.

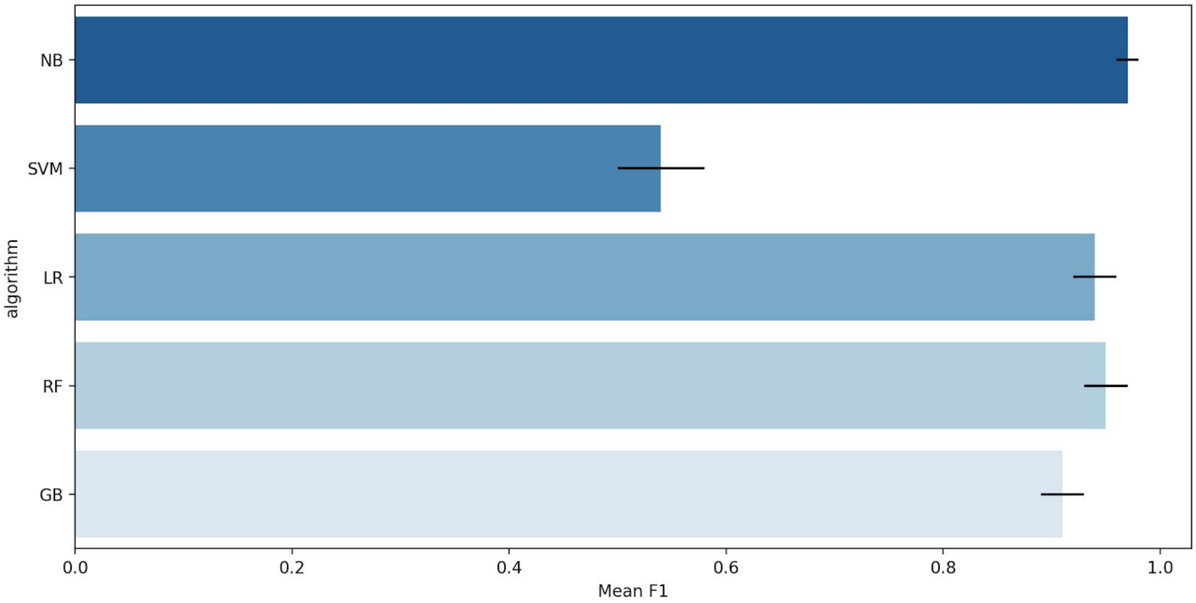
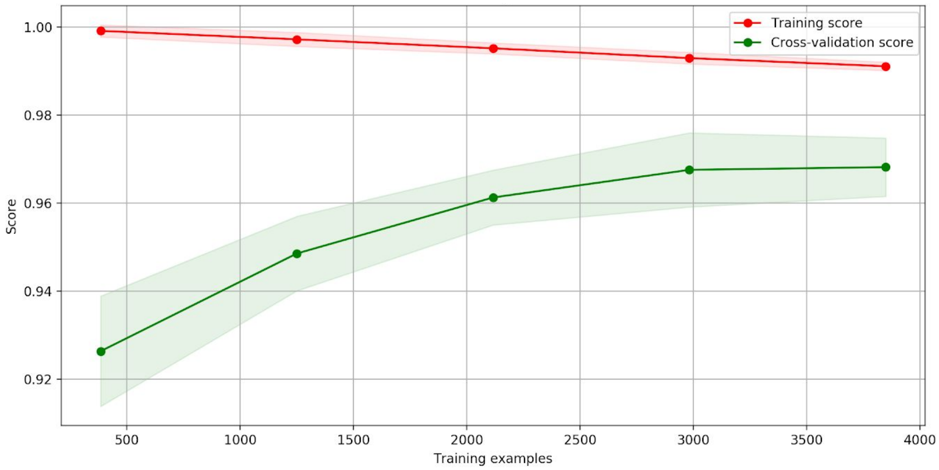
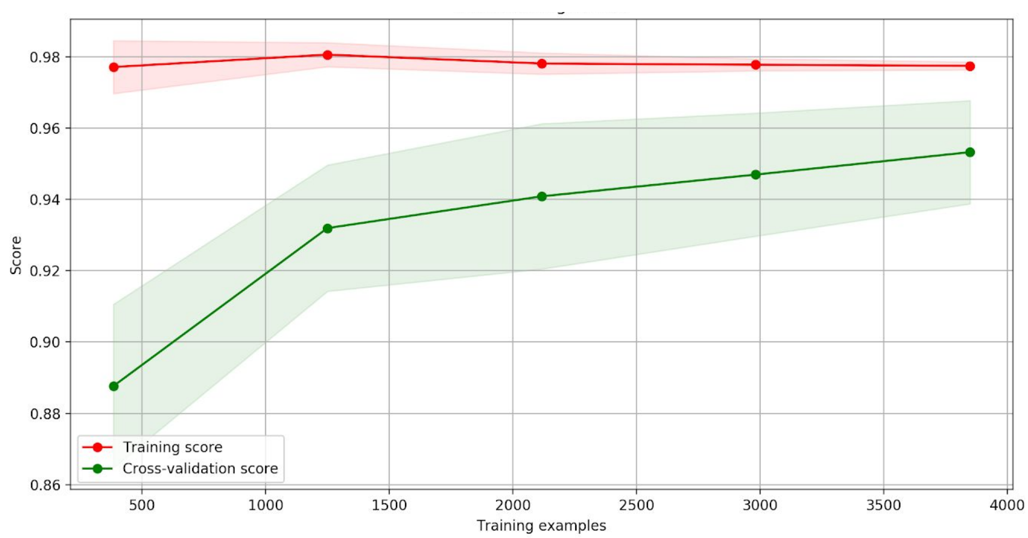


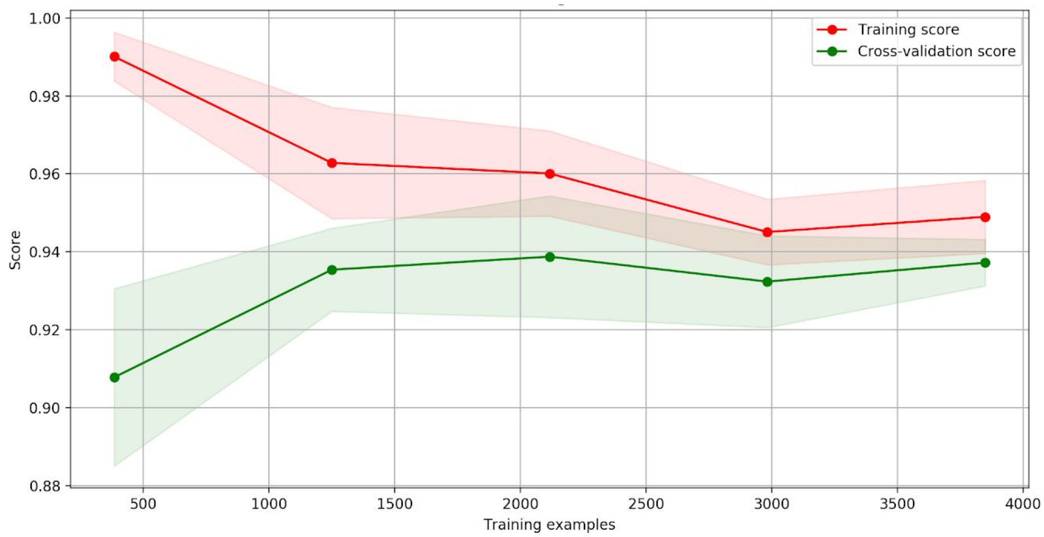
Figure 3. Mean F1 scores for the best performing models of each algorithm. NB=Gaussian Naive-Bayes, SVM=Support Vector Machine, LR=Logistic Regression, RF=Random Forest, GB=Gradient Boosting



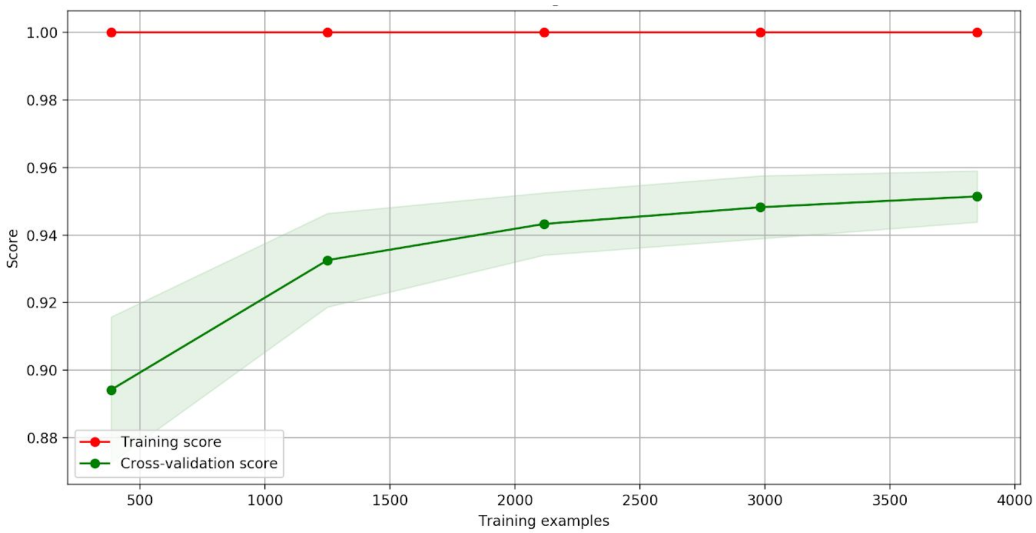
(a) Learning curves of NB model



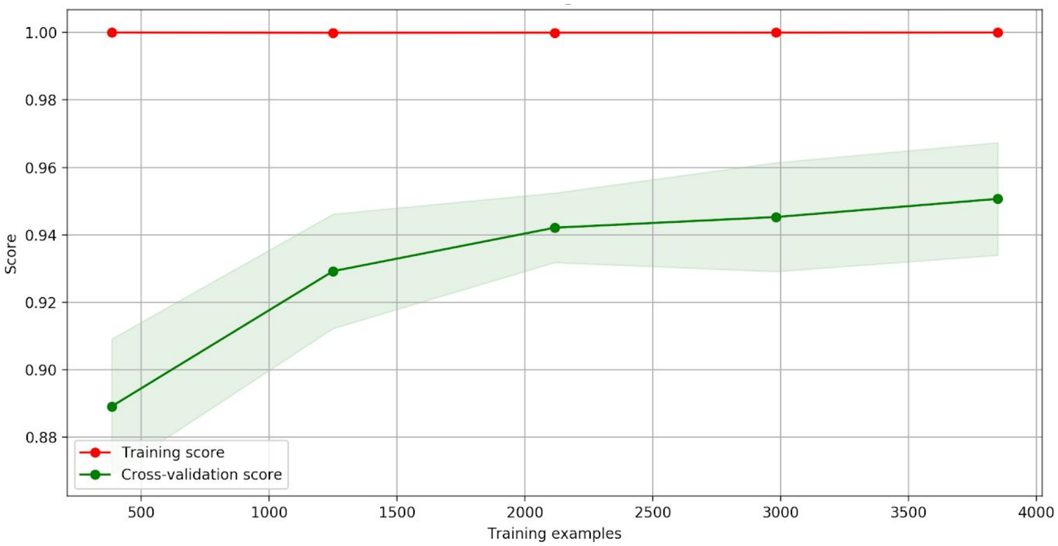
(b) Learning curves of the SVM model



(c) Learning curves of the SVM model



(d) Learning curves of the RF model



(e) Learning curves of the GB model

Figure 4. Learning curves of the trained models

Although NB and LR showed to have pretty good learning results, curves of the LR model converge closer. Therefore, based on this difference in the learning curves of the NB and LR models, the LR model has been selected as the best classifier.

Model Evaluation

The selected LR model has been evaluated using the test set that contains sentences that were transformed using the same configuration of text processing tasks and vectorization of the sentences employed to train the model. The LR model obtained an F1 score of 0.94 in this evaluation. The process followed to build the model has been documented in the jupyter notebook impact classifier.

Baseline Model

As part of the solution, a dictionary-based classifier was implemented and employed as the baseline model. Again descriptions of the social impact of research available in REF were used to construct the dictionary. In particular, I went through the first 230 summaries of impact published here and manually tagged verbs and nouns employed in sentences that contain evidence of the social impact of research. The identified verbs and nouns were extracted and collected in a dictionary. What I discovered during this process is that sentences with evidence of social impact are primarily formed with combinations of these verbs and nouns. The baseline classifier uses regular expressions to find occurrences of terms in the dictionary vocabulary in sentences. The process of building the baseline model has been documented in the jupyter notebook baseline model. The performance of the baseline model was significantly lower than the machine learning classifier, as shown in Figure 5.

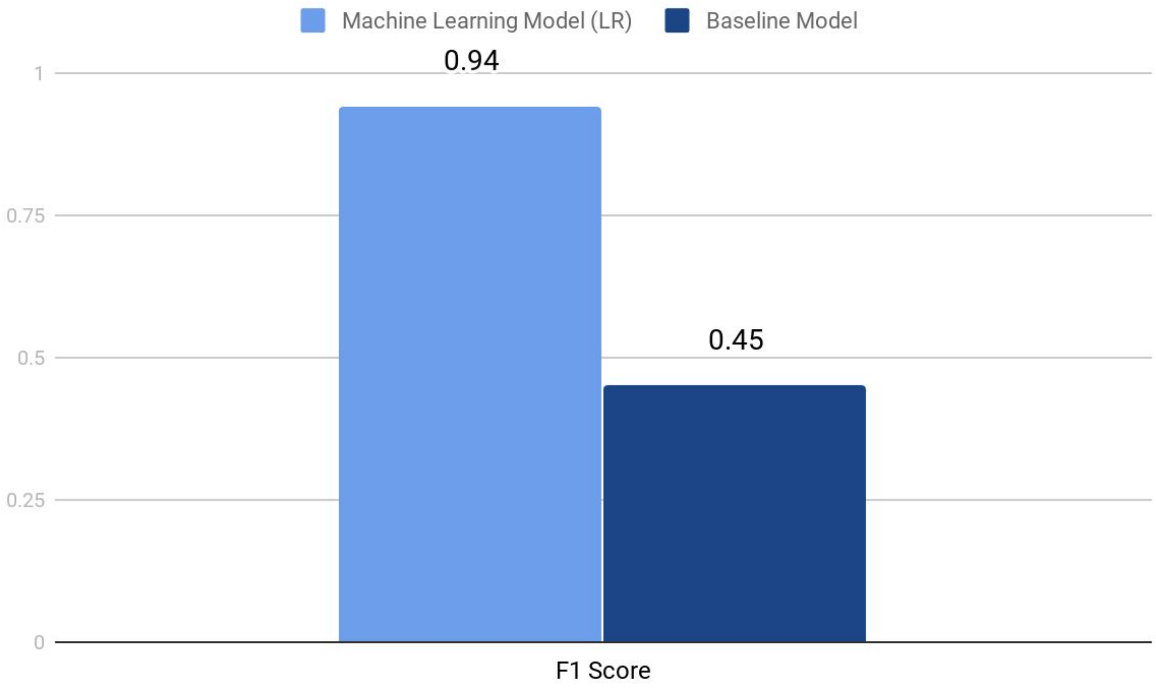
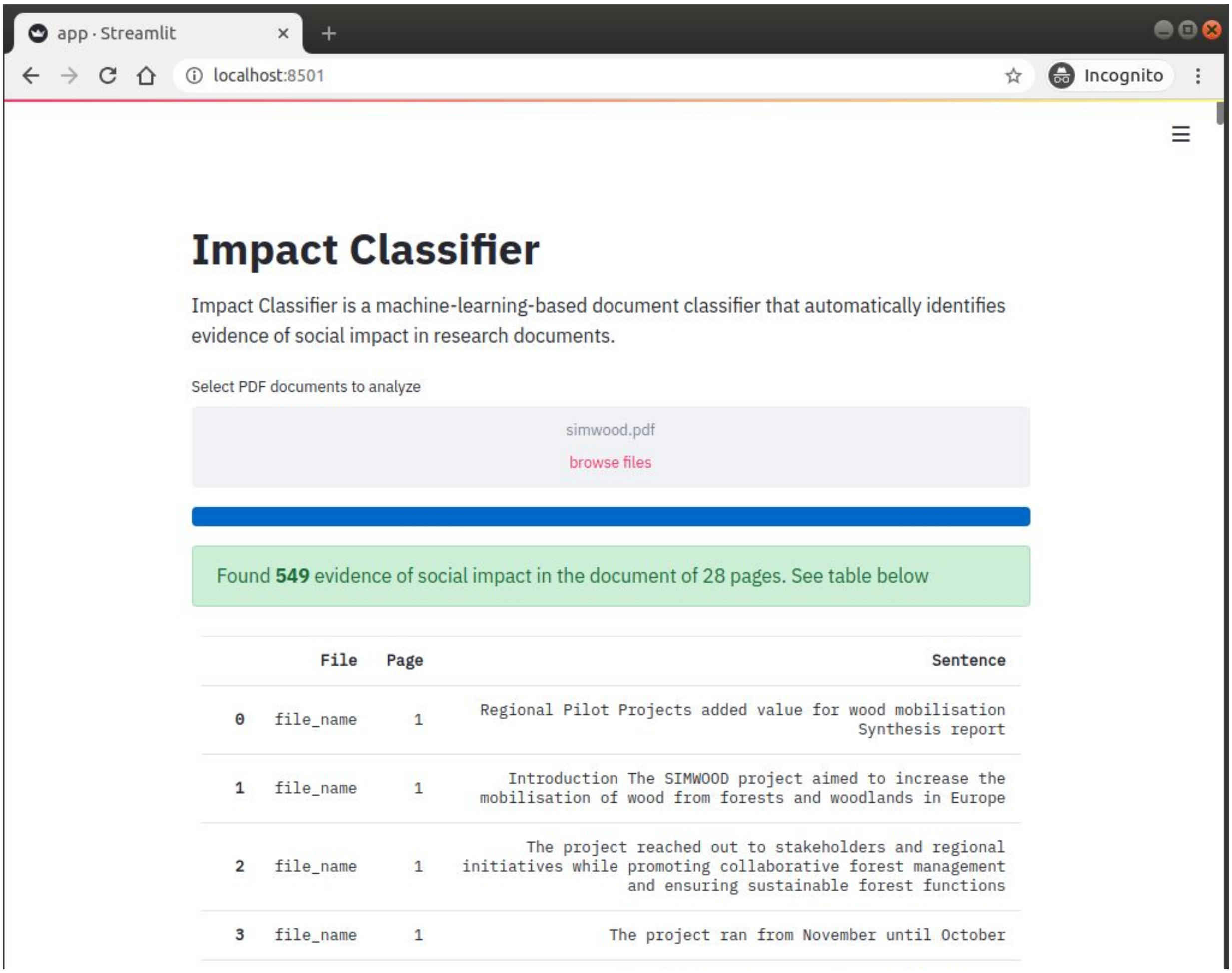


Figure 5. Performance (F1 score) comparison between the machine learning model and the baseline model.

Implementation

The classifier was implemented in a web application using the open-source app framework Stremlit. The application allows users to upload research documents in PDF format. After examining the uploaded document page by page and sentence by sentence, it reports the number of potential evidence of social impact found in the document, as shown in Figure 6. The report informs the page where the evidence was found as well as the sentence that contains the proof. The application was just deployed locally, so at the moment, it is not available online.



Impact Classifier

Impact Classifier is a machine-learning-based document classifier that automatically identifies evidence of social impact in research documents.

Select PDF documents to analyze

simwood.pdf

browse files

Found **549** evidence of social impact in the document of 28 pages. See table below

	File	Page	Sentence
0	file_name	1	Regional Pilot Projects added value for wood mobilisation Synthesis report
1	file_name	1	Introduction The SIMWOOD project aimed to increase the mobilisation of wood from forests and woodlands in Europe
2	file_name	1	The project reached out to stakeholders and regional initiatives while promoting collaborative forest management and ensuring sustainable forest functions
3	file_name	1	The project ran from November until October

Figure 6. Impact classifier, the web application powered by the machine learning classifier and developed to aid in locating evidence of social impact in research documents

References

[1] R. Flecha. Evaluation of the social impact of scientific research. Revista de Fomento Social 73/3–4 (2018), 485–502 485.

[2] A. Shenoy. Text Classification with Extremely Small Datasets. Towards Data Science. Available at https://towardsdatascience.com/text-classification-with-extremely-small-datasets-333d322caee2