Utilizing a Transparency-driven Environment toward Trusted Automatic Genre Classification: A Case Study in Journalism History

Aysenur Bilgin, Laura Hollink,
Jacco van Ossenbruggen
CWI, Amsterdam
{aysenur.bilgin, l.hollink,
jacco.van.ossenbruggen}@cwi.nl

Erik Tjong Kim Sang
Netherlands eScience Center
e.tjongkimsang@esciencecenter.nl

Kim Smeenk, Frank Harbers,
Marcel Broersma
University of Groningen
{k.s.p.smeenk, f.harbers,
m.j.broersma}@rug.nl

Abstract—**With the growing abundance of unlabeled data in real-world tasks, researchers have to rely on the predictions given by black-boxed computational models. However, it is an often neglected fact that these models may be scoring high on accuracy for the wrong reasons. In this paper, we present a practical impact analysis of enabling model transparency by various presentation forms. For this purpose, we developed an environment that empowers non-computer scientists to become practicing data scientists in their own research field. We demonstrate the gradually increasing understanding of journalism historians through a real-world use case study on automatic genre classification of newspaper articles. This study is a first step towards trusted usage of machine learning pipelines in a responsible way.**

I. Introduction

Genre is an important attribute for studying the development of newspapers over time [4] [5] [13]. However, in contrast to topic, information about genre cannot be found using key word search, nor is fine-grained genre information readily available in historical newspaper collections. Therefore, it needs to be added. However, due to the growing size of the data sets in journalism history research [6], manual categorization has become infeasible. In 2017, Harbers and Lonij [14] showed that automatic prediction of genre labels for Dutch newspaper articles is possible with a reasonable labeling accuracy.

From a machine learning perspective, the aim is to improve the accuracy of the prediction model. However, as George Box wrote: "All models are wrong" [3]. Prediction accuracy, albeit carefully evaluated, may not be able to provide a complete assessment of the performance of a classification model, especially if it deals with multiple classes whose distribution over the data sets is unbalanced. On top of this, from a journalism history perspective, knowing what type of errors the classification model makes is crucial for being able to assess whether machine learning model predictions can be trusted. To unite the two perspectives in a single environment, we worked out various types of comprehensible presentations to analyze the decisions of machine learning models. By thus increasing

the transparency of the models, we support the journalism historians in deciding which model's predictions are best for enhancing their research.

As a case study for the transparency-driven environment, we present a real-world scenario using hypotheses from journalism history that can be drawn from applying an automatic method for genre classification of newspaper articles on large-scale unlabeled data. The objective is not only performing well on performance metrics but also being able to explain the predictions of the machine learning pipelines to the journalism historians.

This paper is organized as follows: In Section II, we provide background on transparency in machine learning. Section III presents the design of an environment for transparent machine learning pipelines. The data sets together with the real-world application on journalism history are introduced in Section IV. Next, Section V contains a discussion on the challenges and the lessons learned. Finally, we note some concluding remarks and lay out the future work in Section VI.

II. Background and Related work

The emergence of large open newspaper repositories such as the Dutch Delpher[1] [24] makes millions of newspaper articles available for research. While the abundance of data is fueling data-driven approaches, essentially machine learning, it is also bringing along concerns. Irresponsible data usage and the concealing nature of automation in decision making processes are among these concerns [11] [26]. Initiatives such as the RDS[2] and Explainable AI (XAI) [12] aim to tackle these concerns by increasing awareness on transparency and encouraging the development of new methods to help humans understand the models and appropriately trust them. Furthermore, artificial intelligence, and machine learning in particular, has begun to have progressively more impact in people's lives. In response, regulations concerning transparency have

---

[1] https://delpher.nl/
[2] http://www.responsibledatascience.org/

made their entry to European Union's General Data Protection Regulation (GDPR).[3]

The surge of interest in transparency is tightly coupled with the notion of interpretability. Both concepts have different uses and non-overlapping motivations throughout the literature [17]. According to Lipton [17], transparency is a property of an interpretable model and is examined on three levels: 1) entire model (simulatability), 2) individual components (decomposability), and 3) training algorithm (algorithmic transparency). For the scope of this paper, we consider decomposability to achieve transparency in our analysis as the parameters of the model play an important tangible role in uncovering whether the model's intentions resonate with the theory used for the extrinsic task. Regarding interpretability, we consider the definition provided by Doshi-Velez and Kim [10] and refer to the term as the ability to explain or to present in understandable terms to a human how the machine learning system decides. For the use case study in which we collaborate with journalism historians, it is important to expose for which examples the machine learning decisions are right. Among the approaches for such post-hoc interpretability outlined in Lipton [17], we make use of visualizations (e.g. tabular views) and local explanations (e.g. LIME [21]). There are various open source systems/packages that aim to create, understand, debug, optimize and monitor machine learning pipelines such as TensorBoard[4], Spark ML[5] and Skater[6] [8]. However, the common objective of these architectures is to enrich the understanding of either the developer or the data scientist whereas in our work, the understanding of the journalism historian lies at the core.

III. DESIGN OF AN ENVIRONMENT FOR TRANSPARENT MACHINE LEARNING PIPELINES

In this section, we present the components of an environment that advocates trust and promotes transparency. We have two major objectives for functionality, one is to support the journalism historian by providing detailed insights on the utilization of machine learning pipelines and the other is to facilitate the comparison of performance metrics for the pipelines created using the environment. The machine learning pipelines we consider in this study begin with the data set selection and pre-processing configuration. Presumably, there have been various decisions made while compiling the data sets we use, however, for the scope of this paper, we exclude the details and methods employed for data collection and annotation. Since the data sets relate to the use case study, we provide further details for them in Section IV-C. Furthermore, we note that the environment has the flexibility to integrate a variety of algorithms, text processing tools and techniques, which may not be mentioned in this paper due to time and space limitations.
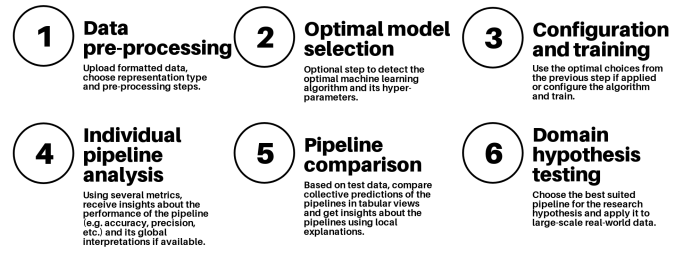

Fig. 1. The workflow of turning data into robust claims in a transparent way. The combination of Step 1 and Step 3 is referred to as ML pipeline.

The workflow of creating and using a machine learning (ML) pipeline in the environment contains six steps as illustrated in Fig. 1 and as described in the following subsections.

A. Step 1: Data pre-processing

We consider three representation types for the documents to be used in the ML pipelines. The representations also include pre-processing of the document as detailed below:

(i) Term Frequency Inverse Document Frequency (TF-IDF): We convert the raw documents to a matrix of TF-IDF scores with the function TfidfVectorizer[7] from the Python package scikit-learn [20]. TF-IDF scores incorporate the frequency of a term in a document (TF) as well as the inverted document frequency (IDF), which is equal to one divided by the number of documents that contain the term. We use four non-default settings for the vectorizer function:
sublinear_tf is set to *True* to use a logarithmic form for frequency, *min_df* is set to *5* as the minimum number of documents a word must be present in to be kept in the vocabulary, *norm* is set to *'l2'* to ensure all our feature vectors have a euclidean norm of 1, and *ngram_range* is set to *(1, 2)* to include both unigrams and bigrams.

(ii) TF-IDF with stop-word removal (TF-IDF (SWR)): As a pre-processing technique, we apply stop-word removal before extracting TF-IDF features from the documents. The stop-word list can be customized according to the use case as further detailed in Section IV-D.

(iii) <NLP suite> with scaling (<NLP suite> (SCL)): By using natural language processing (NLP) tools, we can parse the documents and extract sentences, named entities, part-of-speech tags, etc. that lead to manual curation of various numerical features as further detailed in Section IV-D. It is possible to customize the list of these features as well as selecting which ones to be included in further analysis. When dealing with numerical features, some of the machine learning algorithms may require scaling of the data. For this purpose, we employ RobustScaler[8] from scikit-learn [20], which scales features using methods that are robust to outliers. According to the naming convention

---

[3]https://gdpr-info.eu/art-22-gdpr/
[4]https://www.tensorflow.org/guide/summaries_and_tensorboard
[5]https://spark.apache.org/docs/1.2.2/ml-guide.html
[6]This project is in beta phase and at the time of writing still heavily under development

[7]http://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html
[8]http://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.RobustScaler.html

we use, the name of the NLP suite will be indicated in the brackets. As an example, the representation type will be referred to as FROG (SCL) when we employ Frog[9] as the NLP suite.

B. Step 2: Optimal model selection

This can be considered to be an optional step. Herein, we consider different supervised learning methods to build models of the available data including but not limited to Support Vector Machines, Naive Bayes, and Random Forests.

The optimization is done using scikit-learn's GridSearchCV [10] with the following specified parameters values per machine learning algorithm:

- Support Vector Classifier (SVC) [27]: We opted to include the kernel, C and gamma hyper-parameters in the grid search with the values of 'linear' and 'rbf' for the kernel, [1, 2, 3, 5, 10] for the penalty parameter C, and [0.1, 0.01, 0.001] for gamma, which is the kernel coefficient for 'rbf'.
- Multinomial Naive Bayes (NB) [19]: We decided to include the values [0, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1.0] for the smoothing hyper-parameter alpha.
- Random Forest Classifier (RF) [15]: We chose the Gini impurity (i.e 'gini') for the quality of the split. We varied the number of estimators and the maximum features to be [50, 100, 1000] and [1, 2, 3, 4, 5, 6], consecutively.

Other machine learning algorithms, such as deep learning (multilayer perceptron [23] in Keras[11]), are being considered for addition to the research environment. We used four different scorers[12] (i.e. accuracy, precision (micro), recall (micro) and F1 score (micro)), which are suitable for the multi-class classification problems.

C. Step 3: Configuration and training

This step is composed of three actions: the first one is selecting features when the representation type of the documents allows for it (i.e. see (<NLP suite> (SCL)) representation in Section III-A), the second action is configuring the hyper-parameters of the learning algorithm (that may be suggested in Step 2 or may be carried out as desired by the user), and the third one is training.

The combination of Step 1 and Step 3 is referred to as a machine learning (ML) pipeline.

D. Step 4: Individual pipeline analysis

We make use of visual graphs that can dynamically accommodate multiple classes. One of the most important visualization shown under the pipeline performance is the confusion matrix. We make use of heat map illustration (see Fig. 3) to demonstrate the performance through both 10-fold cross

validation and testing on unseen data. The purpose is to carry out model checking using 10-fold cross validation to expose the potential classification performance. In order to ensure transparency on the pipeline's performance, we train the models on 90% of the data selected for the pipeline and we produce another confusion matrix using the remaining 10% of the data, which is unseen.

Furthermore, we display the numerical values of well-established metrics such as accuracy, (weighted, micro and macro) precision, recall and F1 score resulting from both 10-fold cross validation and testing of the model on unseen data. This step also allows the investigation of global interpretability, which implies knowing about the patterns present in general [10], exclusively for linear algorithms. For this purpose, we show feature importance ranking plots (see Fig. 4 and Fig. 5). For some algorithms (i.e. SVC), the number of these plots depend on the number of classes in the classification task.

E. Step 5: Pipeline comparison

In order to be able to critically evaluate the machine learning pipelines and determine the most suited pipeline for the extrinsic task, we designed three tabular views: 1) explanation view, 2) set-based agreement view, and 3) article-based agreement view. The tabular view is populated using the pipelines in the environment applied to the test sets, which are collocated from the database with duplicates removed. Hence, each document will appear once and will have a unique identifier that can be used to trace in different views. We detail the purpose of these views below:

1) Explanation View: In this view, the main purpose is to communicate the local interpretability, which implies knowing the reasons for a specific labeling decision [10], retrieved using LIME [21]. The view depicts the unique document identifier, the pipeline used for the explanation, the prediction for the classification task using the pipeline and textual or tabular explanation depending on the representation type as part of the pipeline. Further analysis on the local explanations is depicted using visualizations provided by LIME [21] (see Fig. 6). Table I represents the structured information shown in this view.

2) Set-based Agreement View: For a binary classification task, it is convenient to use Venn diagrams when comparing up to three pipelines. However, for multi-class classification tasks, these diagrams would not be as useful. Therefore, we developed a set-based agreement view that can accommodate multiple classes and multiple pipelines to be compared. The view depicts mutually agreeing pipelines on the prediction and the magnitude of the agreement. For data sets that have the ground truth, the documents can be matched against and color-coded with respect to the agreement between the prediction and the ground truth. The <green> documents indicate a correct classification, whereas the {red} documents are misclassified according to the ground truth labels. The documents that belong to unlabeled data sets are colored [blue]. Table II represents the structured information shown in this view.

TABLE I
EXPLANATION VIEW REPRESENTATION

Document #	Pipeline	Prediction	Textual Explanation	Feature Explanation
<A>	<P>	<Class C>	Click here	N/A
<B>	<Q>	<Class B>	N/A	Click here

TABLE II
SET-BASED AGREEMENT VIEW REPRESENTATION WHERE LABEL ACCURACY IS REPRESENTED IN THE DOCUMENTS COLUMN AS: <CORRECT>, {WRONG} AND [UNKNOWN].

Mutually Agreeing Pipelines	Prediction	Number of documents	Documents
<B> <D> <G> <H>	<Class F>	<N>	<A> <K> <L> <M> <N> {P} <T> {U} [W] [Y]

TABLE III
DOCUMENT-BASED AGREEMENT VIEW REPRESENTATION

Document #	Document text	Mutually Agreeing Pipelines	Prediction	True Class
<A>	<Text>	<B> <D> <G> <H>	<Class F>	<Class F>

3) Document-based Agreement View: In the set-based agreement view, the mutually agreeing pipelines serve as index. In this view, however, the documents themselves serve as index. We display mutually agreeing pipelines for the specific document together with the prevailing prediction given by the pipelines and the ground truth if the data set allows for it. Table III represents the structured information shown in this view.

F. Step 6: Domain hypothesis testing

For the domain researchers, the eventual goal for using ML pipelines is to test their hypotheses. In this step, the environment enables the domain researchers to test their hypotheses using plots for a visual analysis. An example of a hypothesis testing plot based on output label frequencies, is given in Fig. 8.

Our machine learning models will produce incorrect output label distributions. Fortunately, we can use the performance of the gold standard data to find out what kind of errors the model makes. This assessment can then be used for re-estimating the label counts on unseen data. The predicted counts will include false positives, which can be removed by multiplying them with the precision score for the relevant label. The predicted counts will miss false negatives, which can be compensated for through dividing them by the recall score for a label. Accordingly, we can re-estimate the output label counts using the following formula:

$$New\_count_{label} = Old\_count_{label} * \frac{Precision_{label}}{Recall_{label}} \quad (1)$$

Note that the precision and recall scores used for re-estimation are only dependent on the output label and not on the input features.

IV. USE-CASE STUDY: AUTOMATIC GENRE CLASSIFICATION IN JOURNALISM HISTORY

This section elaborates on the use-case study to demonstrate the environment presented in Section III. To begin with, we provide background on the automatic genre classification task in journalism history. Next, we present the motivation and hypotheses of this use-case study. Then, we provide the details of the data sets available in the environment and continue with the analysis performed by applying the last two steps of the workflow (i.e. Step 5 and Step 6 of Section III).

A. Automatic Genre Classification in Journalism History

The genre distribution over time gives journalism historians insight into the development of journalism. De Melo & De Assis [18] have proposed a classification model of journalistic genres consisting of five genres: informative, opinionative, interpretative, diversional and utility. These genres are sub-divided into formats, such as the editorial. Their model is based on the communicative purpose of the text. De Haan-Vis & Spooren [9] defined "journalistic prose" as a genre that consists of multiple subgenres, e.g. news reports and interviews. Their genre classification model is based on the (professional) context of the text and the mode of the text. Broersma, Harbers & Den Herder have differentiated between 18 genres, which are distinguished based on their textual form [13]. They have defined the genre labels by looking at how genres have been talked about historically by journalists, how they are defined in journalistic handbooks and self-classified in newspapers. De Melo & De Assis' formats and De Haan-Vis & Spooren's subgenres largely overlap with the genres Broersma, Harbers & Den Herder have distinguished. In this paper, we follow the work of the latter as it enables the domain scientists to get insight into how genres have developed throughout the history. Particularly, we work with the following genres: News (in Dutch: *Nieuwsbericht*), Background (*Achtergrond*), Report (*Verslag*), Interview (*Interview*), Feature (*Reportage*), Op-ed (*Opiniestuk*), Review (*Recensie*), and Column (*Column*).

In 2017, Harbers and Lonij [14] showed that automatic prediction of genre labels for Dutch newspaper articles is possible with a reasonable labeling accuracy. They used the machine learning method Support Vector Machines [2] to build classification models from manually labeled newspaper articles and achieved 65% accuracy. In this study, we refer to the results in Harbers and Lonij [14] as a baseline, and to showcase

	BGS		GS		CGS	
	ALG	HP	ALG	HP	ALG	HP
TF-IDF	SVC	kernel: rbf C:10	SVC	kernel:linear C:3	SVC	kernel:linear C:3
TF-IDF (SWR)	SVC	kernel:linear C:2	SVC	kernel:linear C:3	SVC	kernel:linear C:3
FROG (SCL)	RF	n_estimators:50 max_features:5	RF	n_estimators:1000 max_features:5	RF	n_estimators:1000 max_features:2

the potential value of a transparent environment, we focus on the scientific understanding of the domain scientist.

B. Motivation and Hypotheses

From the end of the 19th century onward, journalism moved from partisan, opinion-oriented journalism to an independent, fact-centered journalism practice [13]. Based on our primary interest in the development of opinionated and fact-centered genres, we look at two hypotheses:

1) The relative amount of *Reports* increased in 1985 compared to 1965.
2) The relative amount of *Features* increased in 1985 compared to 1965.

Like Wevers et al. [28] do for distributions of illustrations, we will create graphs of distributions of the genre labels over time, both for the gold-standard labels and for the predictions of the machine learning pipelines for unlabeled data. We will test if the graphs satisfy these two hypotheses.

C. Data sets

The raw data sets that were made available for this study are detailed below:

- BGS (Balanced Gold Standard)
- GS (Gold Standard)
- CGS (Combined Gold Standard)
- UD (Unlabeled Data)

We assume that the articles in the data sets have been selected randomly and that data selection bias was avoided (but see also Traub et al. [25]). BGS is a balanced data set, which contains 60 articles of each of the eight genres of interest (*Background, Column, Interview, News, Op-ed, Report, Review* and *Feature*). In total, 480 articles were taken from nine different Dutch national newspapers (*Gereformeerd Gezinsblad, Nederlands Dagblad, Algemeen Handelsblad, NRC, Parool, Telegraaf, Volkskrant, Vrije Volk, Waarheid*) and from five years (1955, 1965, 1975, 1985 and 1995). The label distribution over the years is uniform. The manual annotation was performed by a single person.

GS is based on the data set created by Broersma, Harbers & Den Herder [13]. The original data set was a collection of metadata built at a time when no digital versions of the newspapers were available. We used only articles from the years 1965 and 1985 from the newspaper NRC. Not all of

the carefully collected metadata could be linked to the digital articles. The 1,424 articles that could be linked automatically [14] are stored in the Gold Standard (GS) set while the others (884 articles) have been stored in the Unlabeled Data (UD) set. The articles in GS were labeled by various annotators but each article was labeled only by a single annotator. Harbers & Den Herder [13] measured the intercoder agreement for a small part of the data set and found an agreement of 77% (Krippendorf's alpha is 0.67).

CGS is a combination of the two data sets BGS and GS. The order of documents in all of the data sets is randomized. Since BGS, GS, and CGS sets have annotations, we use them both for training and evaluation of the genre classifiers whereas UD is solely used for simulating large-scale real-world data, which would be used for the eventual hypotheses testing.

D. Data pre-processing and feature selection

For document representation type (TF-IDF (SWR)) in which we remove stop-words, we use a modified list of stop-words based on the Dutch stop-words retrieved from the stop-words[13] Python project. The modified list (of 86 words) consists of the original 101 words minus the 15 personal pronouns 'haar', 'hem', 'hij', 'hun', 'ik', 'je', 'me', 'men', 'mij', 'mijn', 'ons', 'u', 'uw', 'ze' and 'zij'. Personal pronouns are excluded from the stop-word list because from a journalism history perspective the distinction between several genres is, among others, based on whether there is a focus on the experience or opinion of the writer. The personal pronoun 'ik' (I) is indicative of this.

Regarding feature selection, we employ structural cues such as syntactic categories [16] (e.g. part-of-speech tags) and use manually curated features based on the definitions of the relevant genres in the codebook for manual content analysis by Broersma, Harbers & Den Herder [14]. These features require parsing and tagging of the text by using NLP tools. For Dutch language, we use an easy-to-use NLP suite known as Frog [1]. The manually curated features extracted from a document are numerical as they represent frequency counts such as the number of adjectives, the number of modal verbs, the number of sentences etc. We also use features such as subjectivity and polarity both of which are retrieved using the module pattern.nl.[14] The total number of manually curated features

---

[13]https://pypi.org/project/stop-words/
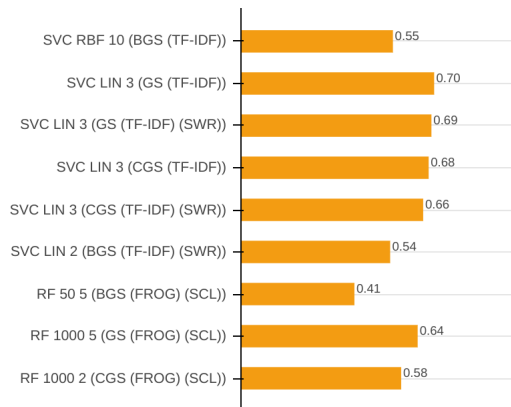[14]https://www.clips.uantwerpen.be/pages/pattern-nl

Fig. 2. Accuracy scores of the ML pipelines available in the environment. The names of the pipelines consist of the algorithm name (RF/SVC) followed by two of its hyper-parameters (50 5/1000 2/1000 5/LIN 2/LIN 3/RBF 10) and then between brackets the training data set (BGS/CGS/GS) and the pre-processing method (FROG/TF-IDF). The use of stop-word removal and scaling is indicated by including SWR or SCL at the end of the pipeline name.

that we consider for this paper is 38.

By the end of the feature extraction and pre-processing steps, we have 9 different training data sets which are listed in Table IV. For the operation of the pipelines, we divide each data set into training (90%) and test (10%) sets using the same random seed so that the test sets for the same data set will contain the same data items.

E. Pipeline comparison

Comparing various pipelines in a systematic way as offered by the transparency-focused environment helps the domain scientist choose the best available one for hypothesis testing. For this study, we chose to investigate 9 different machine learning pipelines that were created using the optimal algorithms and hyper-parameters from Step 2 of the workflow as shown in Table IV.

In the beginning, we observed that the performance measure accuracy[15] was the most appealing selection criterion for the domain scientist when evaluating a machine learning pipeline. From a journalism historian point of view, the pipeline named as SVC LIN 3 (GS (TF-IDF)) looks the most promising (see Fig. 2) followed by the pipeline SVC LIN 3 (GS (TF-IDF) (SWR)), which is trained on the same data set with stop-words removed using the same algorithm and hyper-parameters.

Due to space limitations, we had to limit our further analysis to include three pipelines. In order to choose the candidate pipelines, we make use of the confusion matrices that give further insights on the individual class-level performance. We observe that although the pipeline RF 50 5 (BGS (FROG) (SCL)) scores the lowest accuracy as shown in Fig. 2, the confusion matrix is noted by the journalism historian to be favorable, because it has the most equal distribution on the diagonal compared to others. For the rest of the paper, we

[15]Computed using accuracy_score from scikit-learn (see http://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html)

look at these three ML pipelines whose confusion matrices are illustrated in Fig. 3.

Examining the diagonal of the confusion matrices of the two SVC pipelines in Fig. 3, we see that the classification performance of *News* and *Review* are relatively better than the rest of the genres. However, the performance for the class *Feature* is unfavorable. We interpret this as these models' failure to classify any of the *Feature* articles correctly in the test set. Since one of our hypotheses is focused on *Feature*, it is crucial to observe good performance in this class when determining the most appropriate pipeline for a robust testing of the hypothesis. We note that RF pipeline in Fig. 3 can provide a fair comparison in this regard.

1) Global interpretations: By utilizing the feature importance ranking plots for the SVC pipelines, we can get further insights on what the models deem relevant in general. When we look at the global explanations for *Feature* given by the pipeline SVC LIN 3 (GS (TF-IDF)), we immediately see some interesting patterns. To start with, the recurrence of personal pronouns is interesting and relevant (see Fig. 4). The emphasis on subjective experience in *Feature* is in concordance with the understanding of the genre from a journalism history perspective. Therefore, the fact that the pipeline regards personal pronouns as a distinguished feature is considered to be encouraging. Moreover, the occurrence of personal pronoun 'ik' (I) in *Feature* is identified as indicative in comparison with *Review*, *Background*, *Report* and *News*, but not as indicative in comparison with *Interview*, *Column* and *Op-ed*. This is a favorable pattern as *Interview*, *Column* and *Op-ed* are also genres that focus on either the experience of the journalist or of the subject. The *Column*, for example, focuses even more on the author of the article than the *Feature* does. The features that are found to be important for *Feature* in comparison with these genres are claimed to be certainly interesting as they reinforce the journalism historian's understanding of the genre *Feature* by pointing at the description of a particular event or moment through the usage of words such as 'zaal' (hall; Fig. 4), 'dag' (day; Fig. 5) and 'uur' (hour; Fig. 5).

We observe another relevant pattern with the occurrence of 'gisteren' (yesterday) that can be seen in both of the feature importance ranking plots for *News* (the graph on the right in Fig. 4) and for *Report* (the graph on the right in Fig. 5). According to journalism history literature, these genres both report on newsworthy events in which context the usage of 'gisteren' becomes a meaningful feature. These findings help the journalism historian bridge the gap between their abstract understanding of genres and the concrete features needed to automatically classify them.

In line with our expectations, not all findings are favorable by the journalism historians. However, we note them in Section V-C where we provide a detailed discussion around lessons learned by the help of transparency.

2) Local interpretations: To shed some light on the reasoning of the pipelines, we turn to local interpretations based on the individual articles to improve our understanding.

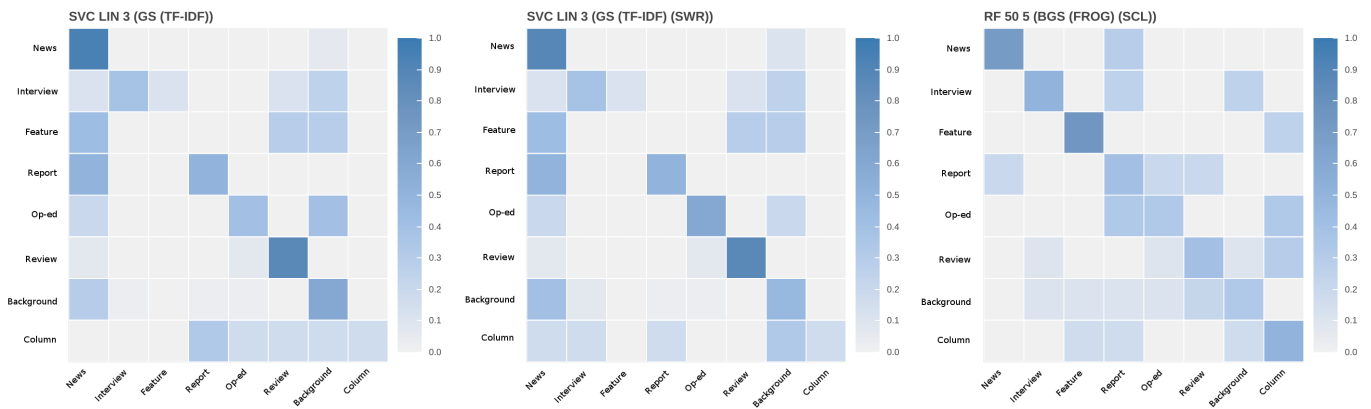First, we look at two articles that are correctly classified by

Fig. 3. Confusion matrices for the selected ML pipelines (populated using only the unseen 10% of the data set)

the three pipelines and compare their explanations in order to determine whether the genre classification is based on relevant features from a journalism history perspective.

- The local explanations given by two pipelines for Article 239, which has a ground truth label as *Feature*, can be seen in Fig. 6. For both explanations (within boxes in Fig. 6), the probability for the top four predictions can be seen on the left and the predictive features for the class that the pipeline has predicted, in this case *Feature*, are listed on the right. The SVC pipeline (left box) shows a focus on the perception of a certain moment, highlighting 'uur' (hour), 'dag' (day), and 'zien' (seeing). Its textual explanations does not show topic words. The feature explanation of RF 50 5 (BGS (FROG) (SCL)) shows some features that can indeed be related to the characteristics of *Feature*. From a journalism history perspective, the large number of adjectives (Fig. 6, right), for example, indicates the focus on the atmosphere and subjective description. The large number of sentences and the number of tokens are also well-known indicators for *Feature*, which tends to have comparatively longer articles together with *Background* and *Interview* in relation to the articles belonging to the other genres.
- The local explanations given by various pipelines for Article 189 has a ground truth label as *Report*. Both SVC pipelines show that the article has been classified as *Report* based on its topic. A large part of the top features having the most importance is related to the act of playing chess: 'zwart' (black), 'verdediging' (defense), 'partij' (game), 'ronde' (round), 'spel' (game) and 'remise' (draw). The explanations given by the pipeline RF 50 5 (BGS (FROG) (SCL)) are more informative and relevant to our understanding of *Report* from a journalism history perspective. We observe that the lack of first and second personal pronouns are strong indicators for the prediction. Since *Report* is an impersonal, factual, chronological description of an event, these characteristics are perceived to be robust features for genre classification.

Next, we look at an article that is misclassified by the SVC

pipelines, which exhibit the highest accuracy scores in the environment.

- Article 175 has a ground truth label as *Interview* and is classified correctly by the pipeline RF 50 5 (BGS (FROG) (SCL)). The local explanations retrieved from LIME for this pipeline show high relevancy with this genre. They include direct quotes, remaining quotation marks, the amount of first person pronouns and (relative) amount of third person pronouns, which are all distinguished characteristics of *Interview*.

  On the other hand, the SVC pipelines misclassified Article 175 and they both predicted it as *Op-ed*. When we look at the local explanations of SVC LIN 3 (GS (TF-IDF)), we observe topic words, such as 'god', 'samenleving' (society), and 'communisten' (communists), which indicate that the pipeline associates political-philosophical topics with the genre *Op-ed* and therefore *Interview* on the same topic is misclassified. From a journalism history perspective, identifying genres through topic is not appealing as topic is not understood as an intrinsic part of genres, while textual form is.

Lastly, we look at an unlabeled article (from the data set UD) on which the three pipelines disagree. Looking at the raw text of the article, we classified Article 60 as *Review*. Both SVC pipelines provide a correct prediction whereas the RF pipeline predicts that Article 60 is a *Background* article. After carefully examining the local explanations of RF 50 5 (BGS (FROG) (SCL)), the prediction seems to be mainly based on the amount of tokens and sentences combined with lack of personal pronouns and lack of exclamation marks. These are indeed relevant characteristics of *Background*. This observation signals that the RF pipeline did not have enough examples to learn the distinction between *Background* and *Review*. Concerning both of the SVC pipelines, the local explanations demonstrate a tendency towards topic classification. From a journalism history perspective, this tendency can be justified due to the fact that *Review* is very closely connected to a specific topic and it can be regarded as an opinion article about cultural products.

de	0.639
van	0.430
in	0.345
gisteren	0.335
het	0.305
de heer	0.274
zal	0.263
amsterdam	0.238
rotterdam	0.235
van onze	0.227
zelf	-0.601
we	-0.605
met een	-0.615
even	-0.617
wat	-0.632
of	-0.634
ze	-0.653
ik	-0.716
maar	-0.728
je	-0.758

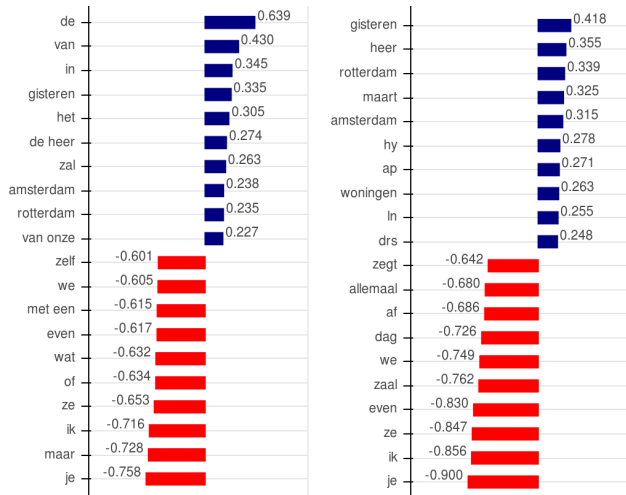| gisteren | 0.418 |
heer	0.355
rotterdam	0.339
maart	0.325
amsterdam	0.315
hy	0.278
ap	0.271
woningen	0.263
ln	0.255
drs	0.248
zegt	-0.642
allemaal	-0.680
af	-0.686
dag	-0.726
we	-0.749
zaal	-0.762
even	-0.830
ze	-0.847
ik	-0.856
je	-0.900

Fig. 4. Feature importance ranking plots comparing the top 10 features found indicative for *News* (blue) in comparison with *Feature* (red). The left graph shows the result of the pipeline SVC LIN 3 (GS (TF-IDF)) which includes stop words while the right graph belongs to the pipeline SVC LIN 3 (GS (TF-IDF) (SWR)) where stop-words are removed. The latter gives more meaningful explanations thereby providing a more solid ground for analysis.

zegt	0.608
dag	0.585
groot	0.497
enkele	0.496
uur	0.467
hem	0.461
hun	0.452
behalve	0.440
nemen	0.437
nacht	0.413
peper	-0.395
telefoon	-0.400
gedicht	-0.401
valt	-0.409
azië	-0.425
keizerskroon	-0.426
hoor	-0.430
postzegel	-0.444
mijn	-0.511
ik	-0.512

we	1.051
zegt	0.763
wij	0.736
ockels	0.725
bezemer	0.700
kelly	0.682
essing	0.676
ploeg	0.666
teheran	0.656
vlijmen	0.622
polen	-0.308
spelers	-0.316
veld	-0.318
ajax	-0.319
streit	-0.330
platonov	-0.374
scheidsrechter	-0.386
rechtbank	-0.399
lubbers	-0.434
gisteren	-0.475

Fig. 5. Feature importance ranking plots given by the pipeline SVC LIN 3 (GS (TF-IDF) (SWR)) comparing the top 10 features found indicative for (left) *Feature* (blue) in comparison with *Column* (red); (right) *Interview* (blue) in comparison with *Report* (red).

F. Domain hypotheses testing

Based on their exploration of the accuracy scores, the confusion matrices, together with the global and local explanations, the domain scientists can make informed decisions regarding which pipeline is better suited for hypotheses testing. The pipeline that can distinguish genre characteristics rather than topic, and that is as accurate as possible would be the best candidate.

After further investigation on also the other pipelines that are among the remaining six (see Section V-B for the details), the pipeline RF 5 50 (BGS (FROG) (SCL)) was chosen to be the best candidate to perform the hypotheses testing since the interpretations based on the features are more convincing, closer to the journalism historians' abstract understanding of the genres and, most importantly, not related to the topic of the articles. Fig. 8 shows the hypothesis graph based on RF 5 50 (BGS (FROG) (SCL)). We observe that the results are compliant with both hypotheses: the amount of *Feature* and *Report* both increase in 1985. However, if we compare Fig. 8 with the genre distribution graph in Fig. 7 based on the manually labeled date by Harbers [13], we see that the RF 5 50 (BGS (FROG) (SCL)) is overestimating the relative amount of *Reports* and *Op-eds* and underestimating the amount of all other genres, especially *Background*. This comparison shows us that even though RF 5 50 (BGS (FROG) (SCL)) is the best pipeline available in terms of feature explanation (Fig. 6) and meaningful genre categorization (Fig. 3), future work is needed to investigate pipelines that are sufficiently reliable and accurate.
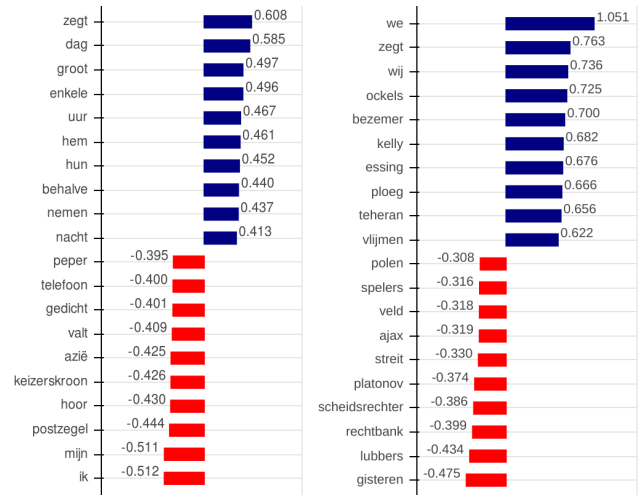
V. DISCUSSION

A. Building a transparent environment

From a computer science perspective, one of the notable challenges of building a transparent environment is the depth of domain relevant knowledge exposure. In other words, how much of the algorithmic details should be shown and to what extent they can be absorbed by the journalism historian is a relevant planning to make during the design process. One way to facilitate the navigation through the details of the data science domain is to provide guiding logic in creation of the pipeline. For example, although it may be trivial to the data scientist not having to select features from a TF-IDF representation, this needs to be explicitly presented to the journalism historian when constructing the pipeline. Thus, it is crucial to address domain related knowledge gaps especially when non-computer scientists are handed the liberty to create their own pipelines for comparative analysis.

During the integration of pre-processing options (e.g. selection of text representation, applying stop-word removal to the text, etc.), we realized that it is important to note the difference between the application of stop-word removal before and after parsing the text with the natural language processing tools as the removal of stop-words may influence the parse-tree. Obviously, this case would be of concern under the circumstances where the representation of the text relies on the natural language processing tools that provide the parse-tree. Similarly as above, the environment needs to communicate this to the journalism historian. We leave the impact analysis of this observation for future work.

The pre-processing of the data, evaluation of classification algorithms and their hyper-parameters can get quite expensive in terms of computational requirements and processing times depending on the number of options considered in the environment. In Step 1 of the workflow, we considered two
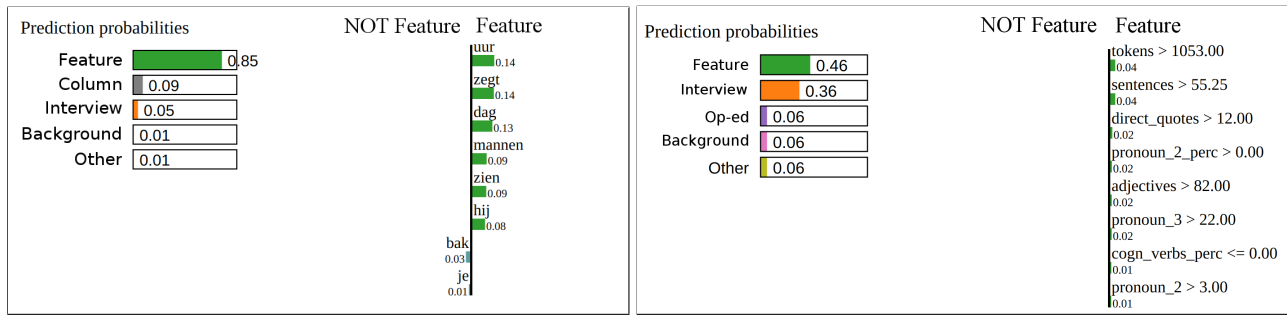
Fig. 6. Local explanations for Article 239 given by the pipeline (left box) SVC LIN 3 (GS (TF-IDF) (SWR)) and (right box) RF 50 5 (BGS (FROG) (SCL)) [Images produced by LIME [21]]

different natural language processing (NLP) tools available for Dutch language, which are SpaCy[16] and Frog. Although we believe that measuring the impact of each step in an ML pipeline is a prerequisite to claim transparency, we leave the impact of using different NLP tools out of scope of this paper. With regards to Steps 2 and 3 in the workflow, it is important to acknowledge the potential impact of various algorithms, parameters and scorers. For example, we looked at 4 scorers during the grid search in Step 2. However, we note that this is not an exhaustive search, but rather a jump-start for the more advantageous numerical performance, which is commonly perceived as the major determinant in selection of an appropriate machine learning pipeline for the journalism historian's task.

B. Choosing the best candidate pipeline: a cyclic process

As noted by the journalism historians, choosing the best candidate pipeline to be employed for hypothesis testing is not straightforward. Although the transparency provided through the global and local explanations lead to some favorable pipelines, the domain scientist still carries doubts about the accuracy scores being low. The explanations reveal the underlying reasoning that the pipelines make, and if the journalism historians do not detect alignment between the journalistic genre definitions and what the ML pipeline has learned, then they will prefer pipelines with lower accuracies.

Having such hesitation ignited a cyclic approach between Steps 4 to 6 in this study. While the determination of the best performing pipeline was at first based on mainly the accuracy score and the confusion matrices, the knowledge gained through navigating in the environment caused the journalism historian to look beyond the accuracy score. To a non-computer scientist, the environment opened new doors to reconsider the impact of the chosen algorithm, the training data set, and the pre-processing on the performance of the pipeline. Under this light, we also looked at other pipelines. Within the remaining six, the pipeline RF 1000 5 (GS (FROG) (SCL)) is worth investigating further as it uses the same representation type as the pipeline RF 50 5 (BGS (FROG) (SCL)) and has an appealing accuracy score. Although the explanations

of RF 50 5 (BGS (FROG) (SCL)) are sometimes generic, the explanations of RF 1000 5 (GS (FROG) (SCL)) in most cases are more specific. For example, the latter has classified Article 239 correctly as *Feature*. The feature explanation graphs show that amount of tokens, third person pronouns, adjectives, sentences, pronouns, remaining quotes, intensifiers and subjectivity are the most important ones. This is a valid explanation as it indicates that *Feature* is distinguished to be a long, subjective article, with relatively many adjectives, third person pronouns, first person pronouns and intensifiers. For journalism historians, these characteristics are parallel to their understanding of *Feature*. The features that they consider crucial such as subjectivity, adjectives, and intensifiers, which are very typical for the subjective description in *Feature*, are as well recognized by the pipeline.

Having had further insight regarding the potential impact of the data set of the pipeline on a local level, it was only tempting to explore this on the global level. For this purpose, we looked at the feature importance rankings given by the pipeline SVC LIN 2 (BGS (TF-IDF) (SWR)). These plots were compared to the global explanations given by the pipeline LIN 3 (GS (TF-IDF)). Subsequently, the particular investigation on the distinction between *Feature* and *Column* was recorded to be entirely different in these two pipelines. The pipeline SVC LIN 2 (BGS (TF-IDF) (SWR)) was found to be showing barely topic-related features, and in fact, it identified *Column* to be about the columnist through the use of first person pronouns.

Going back to the local level for the pipeline SVC LIN 2 (BGS (TF-IDF) (SWR)) was slightly disappointing as a visible tendency for topic classification was found. Article 189, for example, was correctly classified as *Report* by the pipeline. The most important features for the prediction are 'partij' (team), 'hij' (he), 'gisteren' (yesterday), 'zetten' (to move or the moves), 'tilburg' (Tilburg (city name)), 'afgebroken' (cancelled), which are partly related to the topic chess. Although the focus on topic is considerably less clear in comparison with the local explanations given by the pipeline SVC LIN 3 (GS (TD-IDF)), the occurrence of 'partij' (team) and 'zetten' (moves) still signify that predictions are based on topic.

The cyclic approach iteratively led the journalism historians to think that using a pipeline based on the BGS data set and linguistic features is preferable. As a consequence of the
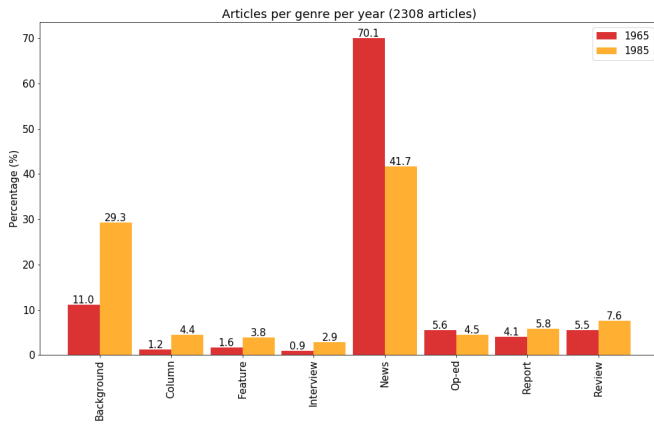
---

[16]https://spacy.io/

Fig. 7. Genre frequency distribution in the Dutch data of Harbers [13]. In comparison with articles in the newspaper NRC from the year 1965, the NRC articles of the year 1985 contained more of the genres *Background*, *Column*, *Feature*, *Interview*, *Report* and *Review*, and less of the genres *News* and *Op-ed*.
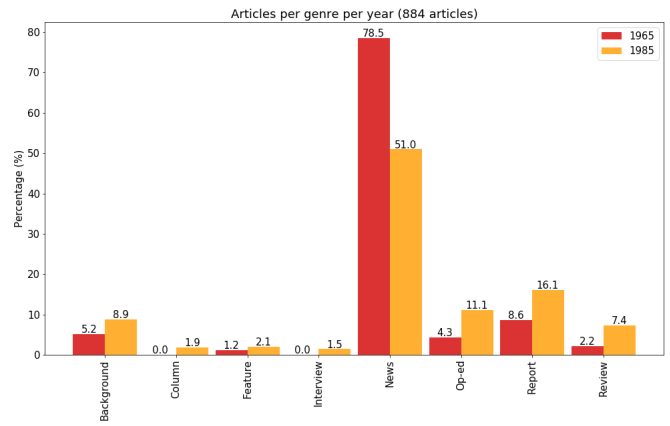
Fig. 8. Genre frequency distribution in the output of the pipeline RF 50 5 (BGS (FROG) (SCL)) for unseen articles of the newspaper NRC. The heights of the bars were recomputed to accommodate for the errors of the machine learning model. There are striking differences with the distributions of *Background*, *Op-ed* and *Report* in Fig. 7 but the two hypotheses given in Section IV-B are satisfied.

transparency support provided by the environment, journalism historians ended up at a totally different point of view where they were freed from the bias of high accuracy scores and they in fact chose the pipeline with the lowest accuracy score to continue their hypotheses testing with. This decision was made based on increased trust and understanding.

C. Lessons learned with the aid of transparency

In journalism history literature, some genre-topic combinations are more frequent than others. For example, *Review* is a genre that always has cultural products as a topic while *Report* is much more often about sport events or court hearings (in the time frame under consideration in this paper). The findings of this study reveal that the pipelines with high accuracy are likely to perform topic classification. This can be visualized in the confusion matrices, especially in those which have darker cells at the diagonal for genres such as *Review* and *Report*.

As postulated in computer science literature, global interpretability implies general patterns between genres. However, our findings do not comply with this statement. For example, the feature importance ranking of *Feature* versus *Column* in Fig. 5, shows that some of the most important features for *Column* are 'peper' (pepper), 'telefoon' (telephone), 'gedicht' (poem), and 'postzegel' (stamp). These words together convey a certain focus on homely and personal event. Although columns often tend to be about these topics, this is not a valid generalization on the genre.

Although on a global level the tendency for topic classification is certainly less visible, on a local level some articles still show how topic plays into the (albeit right) classification. It can be deduced that the data set of the pipeline certainly has an effect, but even a balanced (across genres) data set does not completely alleviate the tendency of the pipelines that use TF-IDF representation to perform classification based on topic rather than genre-related characteristics. On the other hand, the global explanations given by the pipelines that use FROG

representation show favorable compliance with the genre-related characteristics, which are rooted in the journalism history literature. In our study, this is most noted on *Feature* versus all the other genres, where the global explanations are consistently related to the most important characteristics of the *Feature* as a genre.

VI. Concluding remarks and future work

In this paper, we demonstrated the impact of a transparency-led environment that facilitates machine learning pipeline selection for non-computer scientists using a case study on automatic genre classification in journalism history. We noted that transparency-based exploration indeed made a change in the journalism historians' decision making criteria when choosing a machine learning pipeline for hypothesis testing.

Furthermore, we detected the potential to gain valuable insights into how the genre definitions from the journalism history literature can be connected to the relevant features mined from the data sets (see also Broersma [7]). Although claiming to identify genre-related characteristics that are not found in the literature yet is a far-stretching effort at the current maturity of this study, we perceive this as a remarkable advantage in the future of transparency-driven eSciences.

As part of future work, we plan to investigate the impact of diversifying the design decisions within the tools used in each of the steps of the ML pipelines. This includes, for example, trying improved version of LIME that gives higher precision on unseen instances [22], integrating different NLP tools while also applying different NLP configurations, etc. Furthermore, we are interested in the agreement between the collective local explanations (retrieved from the entire data set of the pipeline) and global explanations and whether this will improve the scientific understanding of the domain scientists. Finally, in this paper, we have primarily focused on potential bias and errors in the machine learning models in order to answer our

domain-focused research questions. In the future, we also want to investigate the impact of biases that emerge in the rest of the pipeline such as data selection, data annotation, use of different NLP tools, etc.

REFERENCES

[1] Antal van den Bosch, Bertjan Busser, Sander Canisius, and Walter Daelemans. An efficient memory-based morphosyntactic tagger and parser for dutch. *LOT Occasional Series*, 7:191 – 206, 2007.

[2] Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational Learning Theory (CoLT '92)*, pages 144–152. ACM, Pittsburgh, Pennsylvania, USA, 1992.

[3] George E.P. Box. Science and Statistics. *Journal of the American Statistical Association*, 71(356):791–799, 1976.

[4] Marcel Broersma. An Introduction. In *Form and style in journalism. European newspapers and the representation of news 1880 - 2005*, pages ix–xxix. Peeters Pub & Booksellers, 2007.

[5] Marcel Broersma. The discursive strategy of a subversive genre. In H.W Hoen & M.J. Kemperink, editor, *Vision in Text and Image: The Cultural Turn in the Study of Arts*, volume 30, pages 143 – 158. Peeters Pub & Booksellers, 2008.

[6] Marcel Broersma. Nooit meer bladeren? Digitale krantenarchieven als bron. *Tijdschrift voor Mediageschiedenis (TMG)*, 14(2):29–55, 2011. (In Dutch).

[7] Marcel Broersma. Americanization: or the rhetoric of modernity. How European journalism adapted US norms, practices and conventions. In P. Preston K. Arnold and S. Kinnebrock, editors, *European Communication History Handbook*. New York: Wiley, 2018 (forthcoming).

[8] Pramit Choudhary, Aaron Kramer, and contributors datascience.com team. Skater: Model Interpretation Library, March 2018.

[9] Kirsten de Haan-Vis and Wilbert Spooren. Informalization in dutch journalistic subgenres over time. *Genre in Language, Discourse and Cognition*, 33:137–163, 2016.

[10] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. 2017.

[11] Natasha Duarte, Emma Llanso, and Anna Loup. Mixed Messages? The Limits of Automated Social Media Content Analysis. In *Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*. PMLR, 2018.

[12] David Gunning. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web*, 2017.

[13] Frank Harbers. *Between personal experience and detached information: The development of reporting and the reportage in Great Britain, the Netherlands and France, 1880-2005*. PhD thesis, University of Groningen, 2014.

[14] Frank Harbers and Juliette Lonij. Automating genre classification of historical newspaper articles. Mapping the development of journalisms modes of expression. In *Abstracts DHBenelux 2017 conference (Wednesday 5 July 2017)*. DHBenelux 2017, Utrecht, 2017.

[15] Tin Kam Ho. Random Decision Forests. In *Proceedings of the Third International Conference on Document Analysis and Recognition (ICDAR '95)*, pages 278–282. IEEE, Washington DC.

[16] Brett Kessler, Geoffrey Numberg, and Hinrich Schütze. Automatic detection of text genre. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 32–38. Association for Computational Linguistics, 1997.

[17] Zachary C Lipton. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*, 2016.

[18] José Marques de Melo and Francisco de Assis. Gêneros e formatos jornalísticos: um modelo classificatório. *Intercom-Revista Brasileira de Ciências da Comunicação*, 39(1), 2016.

[19] Marvin Minsky. Steps Toward Artificial Intelligence. *Proceedings of the IRE*, pages 8–30, 1961.

[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[21] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM, 2016.

[22] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *AAAI Conference on Artificial Intelligence*, 2018.

[23] David E. Rumelhart and James L. McClelland. *Parallel Distributed Processing, Volume 1*. MIT Press, 1986.

[24] Thomas Smits. Hidden gems and pointing fingers. *Tijdschrift voor Tijdschriftstudies*, 38:5–7.

[25] Myriam Traub, Thaer Samar, Jacco van Ossenbruggen, Jiyin He, Arjen de Vries, and Lynda Hardman. Querylog-based Assessment of Retrievability Bias in a Large Newspaper Corpus. In *Proceedings of the 16th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)*, pages 7–16, 2016.

[26] Wil M. P. van der Aalst, Martin Bichler, and Armin Heinzl. Responsible data science. *Business & Information Systems Engineering*, 59(5):311–313, Oct 2017.

[27] Vladimir Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.

[28] Melvin Wevers, Thomas Smits, Willem-Jan Faber, and Juliette Lonij. Seeing History: Analyzing Large-scale Historical Visual Datasets Using Deep Neural Networks. In *Abstracts of the 5th DH Benelux Conference (DHBenelux 2018)*, 2018.