A Study of Spoken Audio Processing using Machine Learning for Libraries, Archives and Museums (LAM)

[1]Weijia Xu, [1]Maria Esteva, [2]Peter Cui, [2]Eugene Castillo, [2]Kewen Wang, [3]Hanna-Robbins Hopkins, [3]Tanya Clement, [4]Aaron Choate, [1]Ruizhu Huang

[1] Texas Advanced Computing Center, [2] Department of Computer Science, [3] Department of English, [4] University of Texas Libraries
University of Texas at Austin
Austin, Texas USA

2020 IEEE International Conference on Big Data (Big Data) | 978-1-7281-6251-5/20/$31.00 ©2020 IEEE | DOI: 10.1109/BigData50022.2020.9378438

***Abstract*— As the need to provide access to spoken word audio collections in libraries, archives, and museums (LAM) increases, so does the need to process them efficiently and consistently. Traditionally, audio processing involves listening to the audio files, conducting manual transcription, and applying controlled subject terms to describe them. This workflow takes significant time with each recording. In this study, we investigate if and how machine learning (ML) can facilitate processing of audio collections in a manner that corresponds with LAM best practices. We use the StoryCorps collection of oral histories "Las Historias," and fixed subjects (metadata) that are manually assigned to describe each of them. Our methodology has two main phases. First, audio files are automatically transcribed using two automatic speech recognition (ASR) methods. Next, we build different supervised ML models for label prediction using the transcription data and the existing metadata. Throughout these phases the results are analyzed quantitatively and qualitatively. The workflow is implemented within the flexible web framework IDOLS to lower technical barriers for LAM professionals. By allowing users to submit ML jobs to supercomputers, reproduce workflows, change configurations, and view and provide feedback transparently, this workflow allows users to be in sync with LAM professional values. The study has several outcomes including a comparison of the quality between different transcription methods and the impact of that quality on label prediction accuracy. The study also unveiled the limitations of using manually assigned metadata to build models, to which we suggest alternate strategies for building successful training data.**

Keywords*— *Audio, Machine Learning, shared infrastructure, metadata, libraries, archives, museums, audio transcriptions

I. Introduction

Audio recordings are critical to scientific and cultural inquiries, and providing prompt and accurate access is an imperative for cultural and academic institutions. From performances and oral histories in literary and historical study to scientific sound recordings, the use of audio recordings is increasing exponentially in scope and scale in libraries, archives, and museums (LAM). Meanwhile, LAM professionals are often under-equipped to manage the demands of describing massive audio collections using manual methods [1].

Machine Learning (ML) applications to assist LAM collections management has resulted in both admirers and skeptics. As Ordelman et. al. [2] and others argue [3-6], deploying automation in annotation and collections processing could support LAM professionals charged with preserving and providing access to these materials. For example, ML tools can be used to generate metadata with which users can easily search for items within a collection and that LAM professionals might otherwise have to assign manually [4,5]. While these solutions might improve and accelerate processing, there is a steep road ahead towards researching, developing, and maintaining such methods. As LAM professionals work towards adopting automation, it is critical to question the role and functionality of ML in relation to the unique requirements, best practices, and professional training in this context. To this end, Jakeway et. al point out that ML "*is typically associated with flashy, innovative, transformative, and futuristic problem-solving*" whose operationalization and implementation could be a barrier to entry for LAM professionals who undertake such projects [6]. If LAM concepts, best practices, and values are to be integrated in ML methods for collection descriptions, LAM professionals must be invested in research and development in ML systems.

Our project, AI4AV: Building and Testing Machine Learning Methods for Metadata Generation in Audiovisual Collections (https://hipstas.org/ai4av/), is concerned with creating tools and workflows that are transparent, feasible to use and to share, and that adhere to LAM best practices. Our project, focusing on spoken words audio collections, contributes to ongoing work around the imbrication of ML systems with LAM practices by exploring methods and workflows to support LAM professionals working with digital audio collections. For this, we formed an interdisciplinary team of computer scientists, information scientists, and humanists to identify the roles, technologies, and practices that each can contribute throughout the design, implementation, evaluation, and maintenance of a ML project in the LAM context.

At the beginning of our study we defined the research questions: What are the steps involved in spoken-word audio processing using ML? Can and should we replicate automatically what LAM professionals do manually to describe audio collections? Is metadata produced by LAM professionals

The work presented here has been funded by University of Texas at Austin and National Science Foundation.

using traditional methods adequate to train ML models for classification and description? Is there and if so, What is the impact of the transcription quality in the accuracy of predicted labels? What infrastructure, computational functions, and interfaces are needed within a web framework for LAM professionals to engage with ML tools? Can such a framework be easily configured for different steps and audio collections?

To answer these questions, we designed a methodology that involves automating speech-to-text transcription of audio files and predicting labels that describe their contents for purposes of making them indexable and searchable. Using a proof of concept collection and a prototype web framework, we explored an explainable workflow that combines ML and traditional metadata. The prototype, called IDOLS-AI4AV, is built on supercomputing resources at the Texas Advanced Computing Center at the University of Texas at Austin. Working with 815 audio files from the StoryCorps *Las Historias* collection as a proof of concept audio collection, the workflow leverages existing NSF supported, shared national cyberinfrastructure resources and the framework IDOLS [7-8] as well as open source speech-to-text deep learning tools and ML applications to transcribe audio and to predict labels from the transcripts [9-11].

In this paper we explain the goals of the AI4AV study, review current trends for automated transcription and description of collections, explain our methods and results, and discuss how the findings contribute to traditional descriptive practices in relation to ML applications in LAM.

II. Background

Working with *Historias* collection proved to be a compelling and challenging case study. Founded in 2003, StoryCorps deploys trained facilitators to record and archive short oral histories from people around the United States. Some of the story-collecting initiatives are organized around specific constituency groups and particular events. StoryCorps' *Historias* project captures the *"diverse stories and life experiences"* of the Latino community in the United States. The *Historias* collection documentation states that *Las Historias* will "*ensure that the voices of Latina/Latino people will be preserved and remembered for generations to come*" [12]. The collection of 815 audio files, each of up to one hour of duration, were made available by the partnership between StoryCorps and the Nettie Lee Benson Latin American Collection at the University of Texas at Austin.

To understand the collection's technical provenance our team spoke with StoryCorps staff members to learn how oral histories in the *Historias* collection are created and processed [13]. StoryCorps archivists train facilitators to conduct, record, and catalogue the interviews. Facilitators are self-selected for their interpersonal skills and their desire to tell and learn personal stories rather than their archival background. As a result, the facilitators are a diverse body ranging in age and experience. Full-time facilitators are trained more regularly, while per diem facilitators might only work with StoryCorps sporadically. StoryCorps facilitators introduce participants to the process and permissions associated with recording oral histories and catalogue the interviews in the StoryCorps database during and right after they conduct them. During the recording session, these facilitators take hand-written notes about the interview's content, noting important moments or shifts, and identifying subjects discussed in the beginning, the middle, and end of each interview. StoryCorps does not transcribe the interviews. Thus, the metadata (descriptions and subjects) created at the time of the recording session becomes the metadata that users will later need to find materials in the database.

Because we planned to use the collection's existing metadata to build the ML models, we were especially interested in the process by which subjects and keywords are assigned to describe each interview. *Fixed subjects* are selected from the American Folklife Center's Ethnographic Thesaurus [58]. The terms are updated regularly, and the facilitators can use between five and fifteen terms per recording. On the other hand, *general keywords* can be entered by facilitators when they do not find fixed subjects that adequately describe interview content. In both cases, the emphasis is in describing the general themes of the interviews.

StoryCorps archivists oversee the resultant metadata after it is entered into the database. Due to the intervention of different facilitators and their uneven training and experience, fixed subjects may mean different things to different facilitators, rendering their inconsistent application across interviews. StoryCorps archivists are aware of the shortcomings of the process and are constantly reviewing technologies and methods. And yet, similar shortcomings are common across LAM institutions due to changes in staff members and technologies. Metadata generated for the 815 interviews was provided to our team for use in this study.

III. Related Work

Over the last years, the LAM community has been addressing research and development of AI methods for digital libraries and archives materials processing [14-16]. Transcribing speech to text using AI methods can enhance and accelerate access and research to large collections, and many projects are focused on obtaining good transcriptions automatically [17-18]. Because the main goal of our study is to predict subjects that describe the content of spoken-word audio, our interest is in exploring whether the quality of the transcription affects label prediction accuracy.

Like ours, other ML projects take advantage of the trove of metadata produced through manual cataloguing. Annif is an open source microservice to automatically assign subject headings to textual materials [19]. An interesting feature of this project is its interface that allows easy submission of texts, selection of ML methods, results evaluation, and feedback to the model. To train its models, Annif uses manually assigned subject headings, and terms from controlled vocabularies. This very complete and thoughtful project considers human cataloguing the gold standard for comparing results. While in our project we also use human-selected fixed subjects as

training data, we suggest that depending on how the subjects are assigned they may introduce inconsistencies or biases to the models. For example, at the University of Utah Libraries, an image indexing feasibility study using off-the-shelf AI software and human-made metadata for training, found that digital library metadata may need to be re-designed for ML applications [20].

The Collections as Data project encompasses resources and proof of concepts focused on using computational methods to process digital data collected in cultural institutions in consideration with the values that characterize their practices and services [21]. In alignment with the project's Santa Barbara principles, our prototype ML framework aims to lower the barriers to computational use of digital collections, addresses analysis of both data and metadata, and can be shared by staff members and by many institutions [22]. A unique contribution of our study is that the statistical exploration of human-made metadata for use as training data, allowed us to better understand traditional metadata practices in LAM and make suggestions towards improving them for purposes of building ML models.

Automated speech recognition (ASR) is an active research field with a long history. The basic goal of ASR is to map audio input into an acoustic model representation which is joined to a language model for generating written transcripts as outputs. A common approach assumes a probabilistic model between the acoustic representation and the language representation so that the decoded string has the maximum a posteriori (MAP) probability [23, 24]. To this end, a number of models have been proposed over the years including: hidden Markov chain model [25], template-based approach [26], stochastic modeling [27], and vector quantization (VQ) [28]. There are numerous advances in speech recognition using neural networks [29-31]. Examples include feed-forward neural network acoustic models [32,33], convolutional networks [34,35], and recurrent neural networks [36-39]. The DeepSpeech tool, used in this study, is an end-to-end pipeline for ASR that combines deep neural networks with convolutional layers [9,10].

Text classification remains a problem in assigning predefined classes (labels) to an unlabeled text document [40]. The general solution includes representing the text in a vector space or a graph model with selected features such as word grams, topics, taxonomies, or metadata [41-43]. Once the text is well represented, supervised machine learning methods such as support vector machine, Naïve Bayes, and decision trees are used to build classification models [44-47]. With increased data availability, deep neural networks applications such as BERT, ALBERT, and GPT have been found successful for text classification [11, 49, 50]. For text classification methods to infer labels for an input document, the labels must be from a finite set of predefined categories from which enough documents are available to train the classification model. In our study, we use fixed subjects applied by humans as categorical labels for each transcript/audio file. As a result, our ML model will classify the individual interviews based on how the fixed subjects were assigned in the first place.

IV. Methodology

With the goal of generating labels to describe audio files automatically, this study explored different state-of-the-art ASR and text classification methods. In the process we assessed if the quality of the transcriptions influences label prediction results and evaluated the use of manually assigned metadata as training data. For purposes of this study we focused on fixed subject metadata.

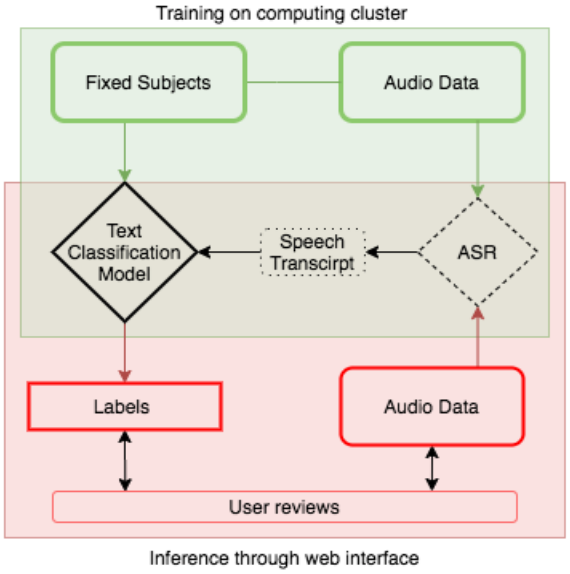

Fig. 1 Methodology workflow overview.

Fig. 1 shows an overview of the methodology. First, we converted the audio interviews to texts. For this, we used two different ASR tools for automatic speech-to-text transcription, and compared the results against human-made transcriptions as ground truth. Following, we conducted statistical analyses of the fixed subjects applied to the interviews by the facilitators. The results of the analyses informed the design of the experiments to predict labels. Once the models are built, it can infer labels for new audio input. As we refined the methodology into audio processing workflow steps, we were designing configurable web interfaces to facilitate conducting the tasks in shareable, flexible, and transparent ways.

Throughout the study we also carried out qualitative review of the results to verify the outputs in relation to what users would read and search for. We also conducted exercises to learn how fixed subjects are manually applied and to understand how this practice influences the predicted labels results. Indeed, the qualitative observations suggested ways to improve label implementation as training data.

A. Speech to Text Conversion

815 audio files from the *Historias* collection were transcribed to text using DeepSpeech (DS) version 0.8.2 (https://github.com/mozilla/DeepSpeech) and 311 where transcribed using Google's Speech-to-Text (GST) service

(https://cloud.google.com/speech-to-text). We also obtained 81 human transcriptions (HT) through the University of Texas Libraries Captioning and Transcription Services [51]. The difference between the number of transcriptions obtained for each method is due to budgetary constraints.

DS is an open-source automatic speech recognition tool that uses a pre-trained English model to transcribe spoken audio [9,10, 52] . While some interviews in *Historias* include words in Spanish, we only used the freely available pre-trained English model implemented with Tensorflow [53]. According to its release documentation, the pre-trained acoustic model was trained on American English reporting 5.97%-word error rate (WER)_on the LibriSpeech test corpus [54]. Due to limitations of its training data, the model has biases towards high quality recordings with minimum noises, and to speech from US male accents.

We also experimented with GST cloud commercial service [55]. The standard model usage is about 2.4 cents per minute. This rate is in addition to other computing costs such as cloud storage and virtual machine resources if applicable. This model is more complete than DS. It supports 125 languages and has domain specific models to choose from. The service also has optimization for audio recordings with low sampling rates, and post processing steps to convert numbers, addresses, and times.

While the Historias collection has words and names in Spanish, to simplify the analyses we only used English language trained ASR models. We here note that there is a difference between transcribing texts in different languages and transcribing mixed language texts, the latter being a more complex problem to solve.

The quality of the transcriptions derived from DS and GST was evaluated qualitatively by members of the team who listened to a sample of audio files and read the corresponding transcriptions. The quantitative analysis was conducted against the HT using word error rate (WER). WER is the percentage of mismatched words between a transcribed text and the ground truth text. Results are reported in Section V.

B. Text Classification - Models for Label Prediction

We investigated how to utilize ML methods to predict labels based on the interviews' transcriptions and the fixed subjects assigned by the facilitators as metadata. Because each interview is associated with a variable number (5 to 17) of very diverse fixed subjects (Figure 2), to build the model we decided to limit the number and use the top-20 (Figure 3).

To predict multiple labels from a transcript, we explored two supervised learning approaches, top-N and multi-Label. In the top-N approach, all fixed subjects associated with one interview are first ranked based on their overall frequency across all the interviews, and only the top-n are selected for training the model. Consequently, the resultant ML model will only predict n labels. For example, when n is one, one label from each interview is selected and the model will predict just one label per interview. In the multi-label approach, there is no ranking nor preselection of fixed subjects. Instead, the ML model is built to infer a vector of likelihood; this is how likely a label is associated with an interview. Combining different approaches and ML tools we developed and evaluated the three models detailed below.

The first model, hereafter referred to as ***RF***, utilizes Random Forest learning methods. In this model, each transcript is represented as word frequency vectors based on the bag-of-words model. From the transcripts and the selected fixed subjects, the RF model is trained to infer if a given label belongs to the top-n fixed subjects or not. The results presented here are from cases where n is 3.

The second model, hereafter referred to as ***DL-TopN***, is built to make similar inferences for top-n labels using a deep learning (DL) long-short-term memory neural network within the BERT library. We used the word2vec from BERT to convert each transcript to a vector representation. Similarly, to the RF model, we focused on cases where n is 3.

The third model, hereafter referred to as ***DL-multi***, is built to infer how likely an input transcription can be associated with each of the selected 20 fixed subjects. Multiple fixed subjects assigned to an interview are represented in a twenty-dimensional vector, each vector corresponding to one fixed subject. In the training data, this vector is binary, such that the value of each dimension is either 1, indicating that the subject is assigned, or 0 indicating otherwise. From this vector, top-n labels are dynamically computed and are not restricted to a preselected value. For the testing data, the model computes a similar 20-dimensional vector in which, the value of each dimension indicates the likelihood of the association between a predicted label and the transcript. To build this multi-classification model we used the BERT library. In all cases, the results were evaluated based on accuracy, which is the percentage of predicted labels from the original set of fixed subjects per interview.

C. Selecting Subsets of Fixed Subjects

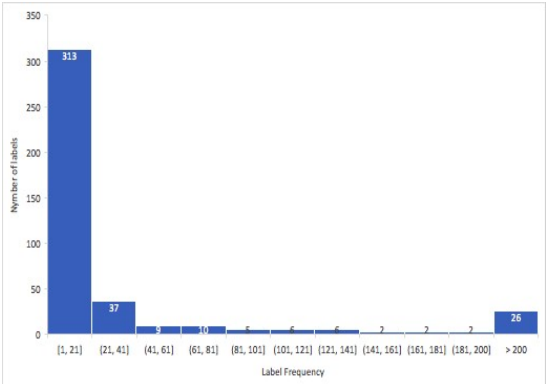

Fig. 2 Histogram of number of fixed subjects over number of occurrences in the collection.

Building a balanced model to predict labels per interview implies that the training data is representative of the content of all the interviews. Towards that end, conducting statistical analyses to learn the characteristics and fitness of the data that

will be used to train the model is a requisite in all ML applications. Such analyses evaluate data completeness, balance, and coverage, to anticipate possible biases in the results [56].

Fig. 2 shows an analysis of frequency of usage of all fixed subjects appearing in the *Historias* collection. The facilitators used a total of 418 unique fixed subjects to describe the *Historias* collection. Despite the limit of 15 subjects suggested by StoryCorps, each interview is associated with 5 to ~25 fixed subjects, and there are significant variations in their occurrence. For example, 313 out of 418 fixed subjects are used in less than 20 interviews, and only 26 are used more than 200 times.

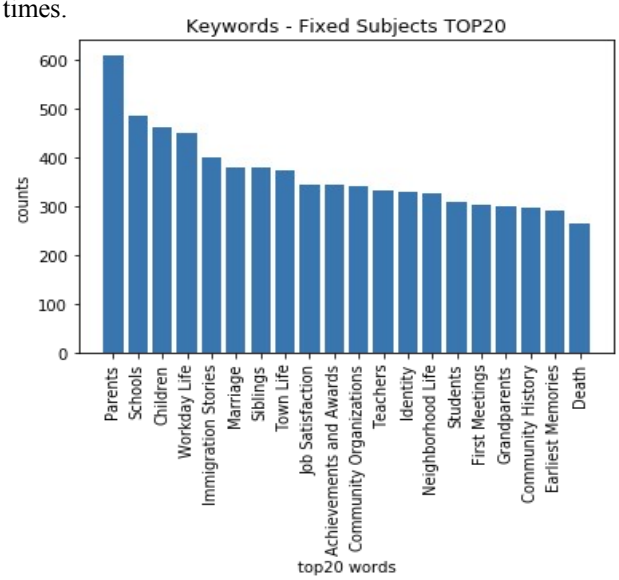

Fig. 3 Distribution of the top 20 fixed subjects across number of interviews.

To build a classification model for label prediction, we needed to select a subset of fixed subjects with enough number of occurrences in the collection. Fig. 3 above shows the number of times that each fixed subject was used to describe the interviews. After observing their frequency distribution across the entire collection, we decided to focus on the top 20 most frequent as training data.

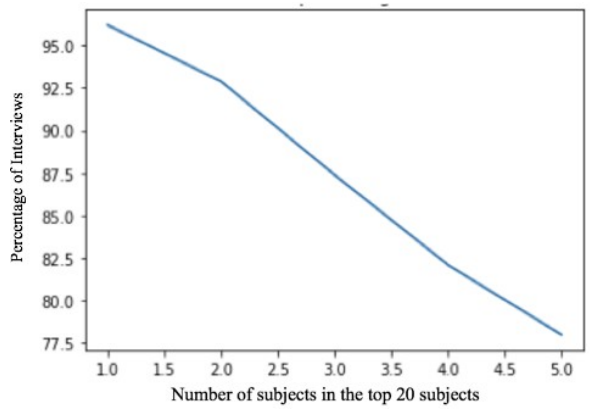

Fig. 4. Percentage of interviews with top-20 selected subjects.

We also needed to decide how many labels we could predict with confidence. For this we assessed coverage, understood as the percentage in which any of the top 20 fixed subjects appear per interview. Fig. 4 shows that each interview was assigned at least one fixed subject from the top 20. More than three top-20 fixed subjects were assigned to less than 90% of the interviews, and more than five to 78% of the interviews. The results suggested that we could predict at least 3 labels with confidence from this ranked list of 20.

D. Web Interface for Interactive Access

IDOLS (idols.tacc.utexas.edu) is a web-based API platform developed with support from the National Science Foundation [7,8]. In this project we use it as the gateway to Machine Learning and Natural Language Processing tools installed in High Performance Computing (HPC) resources at the Texas Advanced Computing Center. IDOLS bridges applications - in this case for purposes of audio transcription and keyword prediction - with remote compute resources. Its goal is to provide a low barrier to increase HPC adoption through interactive interfaces. The IDOLS framework enables creating and customizing web applications from a configuration file. The web application is self-contained and can be deployed without alleviated system privilege. Therefore, ad-hoc analysis routines can be described and preserved in a format that can be shared and re-used. The application can also be preserved through the configuration file for reproducibility. Utilizing IDOLS, we have developed an on-demand web application that can run on remote computing clusters. In this study, the application was configured for the specific needs of the AI4AV project. The different interfaces allow accessing raw audio files, running speech transcription processes, saving and editing the code to configure the workflow and its interface, obtaining predicted labels, and providing feedback to the model. These capabilities were designed to add transparency and reproducibility to the professional process values within LAM.

V. Results

A. Speech to Text: Quality Comparisons

To evaluate the quality of the ASR methods we compared the transcriptions generated by HT to the outputs of DS, and GST, both quantitatively and qualitatively. Qualitatively, by reading the texts, the team concluded that the HT are the most comprehensible, as both in the DS and in the GST outputs many transcribed words are gibberish without meaning. The human transcriber added punctuation marks and annotated the different speakers in the interviews. Instead, the ASR models generate text outputs without punctuation or distinctions between speakers. And yet, while both the DS and GST transcripts are hard for readers to follow, the latter are more understandable than the former. We were interested in learning if and comprehensible texts influence label prediction accuracy.

To quantify the quality of the ASR results we used the HT as ground truth and compared those to DS and GST outputs. The common evaluation metric, word error rate (WER), was calculated for each comparison [59,60]. The WER first aligns transcribed sentences together and then computes the number

of changes required to transform one sequence to the other sequence. The word error rate is defined as:

$$WER = \frac{!"\#\$\%\&\quad ()\quad *(*+,\quad \%-.*/\quad \&\%0".\&\%}{*(*+,\quad ,\%!12*\quad ()\quad *2\%\quad \&\%)\%\&\%!3\%\quad /\%0"\%!3\%}$$

For this quality assessment we used 27 sets that included the three types of transcriptions. Given that the number of transcriptions per method is different, the number 27 is related to the availability of the same audio transcribed with the three methods. Before comparison each transcript was processed including stop words removal, punctuation marks removal, stemming, and lemmatization [57]. Following, each transcript was converted to a list of tokens whose numbers ranged from 1207 to 3037.

Fig. 5 shows the WER results plotted in relation to the number of tokens found in the HT. The average WER for DS and GST are 0.67 and 0.47 respectively. Note that the WER curves from both methods are consistent and seem independent of the length of the transcribed file. The consistent lower error rate for the GST model implies that it is better than the DS one. While GST has a lower error rate than DS, both results point to significant challenges in ASR.

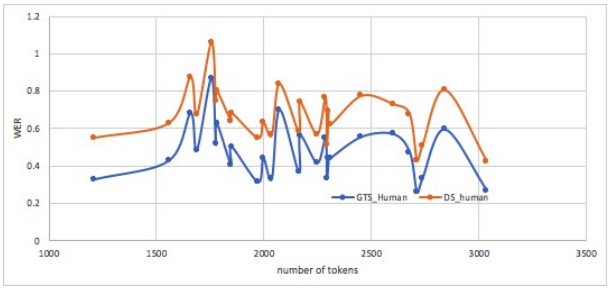

Fig. 5 WER for DS vs HT (red) and GST vs HT (blue) comparisons for transcripts of different length.

Figure 6 below shows the comparisons between three sets of transcription results: DS vs. HT (red), GST vs. HT (blue) and GST vs. DS (gray). The average WER between GST and DS is 0.49 which is better than the WER between DS and HT. This suggests some commonality between the DS and GST models. Between the two ASR methods, GST gives better results than DS. However, neither are close to the quality of the HT outputs. These results coincide with our qualitative assessment.

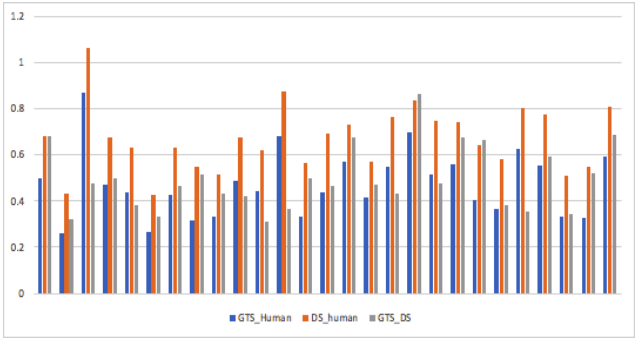

Fig 6 WER results comparisons between DS, GST, and HT.

C. Text to Label Results

We mentioned that we were interested in exploring the use of ASR outputs to predict descriptive labels and to assess the implications of the transcriptions' quality on the predicted labels' accuracy. To explore this question, we conducted three experiments.

The first experiment compared how well the three models described in section B, can predict the top-3 labels using transcripts generated by DS. Table I shows the results. Each model was trained using 80% randomly selected DS transcripts, and tested over the remaining 20%. For each transcript, the average accuracy was computed as the percentage of correct labels over all predicted labels. Results are shown in Table 1. The DL-TopN model yields the best results, and the accuracy of the RF model is a close second. The DL-multi model has the lowest accuracy score, which is 7% worse than the DL-TopN model. The results of this experiment indicate the viability of label prediction even with low quality transcriptions. The comparisons do not show a significant advantage between the two DL-based models over the RF model, a result that may be impacted by two factors: the quality of the transcriptions, and the amount of data available for training and testing.

Table I Average accuracy of top-3 label predictions for the three models using transcripts generated DS.

DS transcripts	RF Model	DL-TopN	DL-multi
Average Accuracy	0.70	0.71	0.66

Once it was established that we could predict labels based on the least intelligible ASR output, in the second experiment we further investigated the relation between the transcription quality and the label prediction accuracy. For this we created three test sets based on their transcription method. As we mentioned, due to budget restrictions, the number of transcripts tested for each method was different, i.e. 815, 311 and 81 for DS, GST, and HT respectively.

Table II. Average accuracy of the top-3 label prediction using different transcription methods and classification models.

Average accuracy	**DS**	**GST**	**HT**
Number of transcripts	815	311	81
DL-TopN	0.71	0.67	0.66
DL-Multi	0.66	0.96	0.85

Table II shows the average accuracy results for the top-3 labels using DL-TopN and DL-Multi model, which rendered the best results in the first experiment, over the three sets of transcripts. The HT has the least number of transcripts available. For the DL-TopN model, the average accuracy over the top-3 labels is similar between the three sets of transcripts.

This implies that this model is less sensitive to the quality of the transcripts. Further investigation showed that the model usually discovers the dominate labels but it lacks specificity. The DL-Multi model achieved better results testing on the higher quality transcripts GST and HT. These results indicate that the DL-Multi model is more sensitive to the quality of transcripts and in addition to the general trend, it can pick up fine differences among classes/labels. It was a surprise to see that the model built with GST transcripts outperforms HT. However, one possible reason could be that the lack of a larger set of HT data makes the model less reliable. The results incited us to further test the DL-Multi model using combinations of training and testing data.

The results of this experiment were computed for different training/testing transcription sets. HT GST and DS are results from using HT, GST and DS respectively for both training and testing. HT* denotates results from the model trained with HT and tested against DS. GST* denotates results from the model trained with GST and tested on DS. Label prediction accuracy degradation is observed with the DL-Multi model when predicting more than three labels. Fig.7 above shows how the average top-n labels accuracy decreases as the number n increases. The best results are found for models trained and tested on GST with 96% accuracy for prediction of the top-3 labels, and for up to top-12 labels with acceptable accuracy (0.80). In comparison, the model trained and tested on HT has an accuracy of 0.85, and the model trained and tested on DS has 0.66 accuracy. When testing transcripts from DS (DS, HT* and GST*), all three models show lower accuracy values and are similar to each other around 0.66. These results suggest that the best model for label prediction is DL-Multi which renders the best tradeoff between accuracy and transcription method quality.

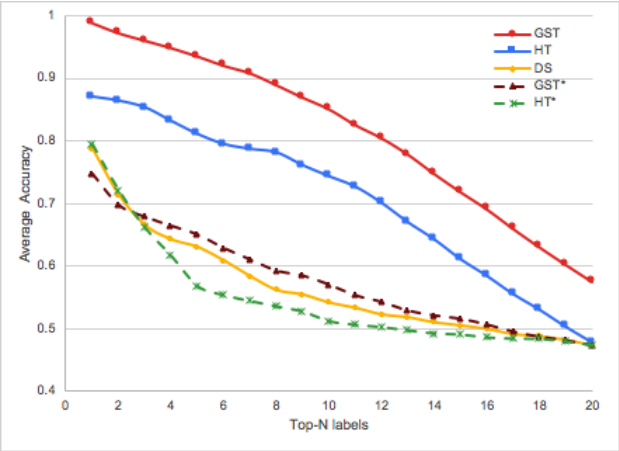

Fig. 7 Average accuracy comparisons for DL-Multi model with different sets of transcripts.

We also reviewed the results qualitatively. For a sample of interviews, we read the labels generated with the DL-Multi model tested on each transcript method. We noted that the labels predicted using DS transcripts are exactly the same across all the interviews. This observation confirmed that the quality of the transcription is relevant to distinguish unique and diverse classes/labels. In turn, this agrees with the quantitative results. Otherwise the models converge to the dominant trends in the collection and thus produce same labels for all transcriptions. And yet, while text-to-label prediction is improved with high quality ASR transcriptions such as GST, predicting labels from low quality transcriptions such as those generated through DS should not be dismissed. Especially if considering that DS is open source application that can be further improved. Specially for LAM, if well trained, it can be used to process large amounts of data at lower costs.

VI. Expert Assessment and Web Interface

A. Expert Assessment of Metadata as Training Data

The involvement of LAM professionals to design, steer, and evaluate the development and results of ML projects is key to produce outcomes in line with best practices and values in the space. The quantitative results of this study indicate that it is possible to predict labels from existing metadata with good to reasonable accuracy. However, the predicted labels will be as good as the metadata selected by LAM professionals. Thus, we sought to understand the adequacy of manual metadata generated by LAM to train ML models.

Learning about the mechanics of how fixed subjects are applied in the *Historias* project (Section II), was the first step towards exploring this question. In addition, the distribution and coverage of fixed subjects (Figures 3 and 4), provided quantitative insights on the metadata practices. We learned which labels were used the most as well as the consistency of their application across the interviews.

When our team reviewed the top 20 fixed subjects that would be used to train the ML model (Figure 3), we discussed if they represented the diversity of content of the interviews. Considering their purpose of describing general themes across all the interviews, many seemed overlapping or redundant. For example, a cluster of fixed subjects: students, teachers, and schools, could be consolidated into one encompassing subject such as education or schooling, and the same could be said about community organizations and community history.

To better understand fixed subjects' selection, their granularity, and their consistency, we devised two simple exercises. In the first one, we aimed to identify if expert curation could render a more diverse short list of fixed subjects for training than the one obtained by ranking. Four members of our team set out to consolidate a list of top 50 fixed subjects into 20 by classifying related terms into broader ones. We later met as a group to agree on a final list. When we compared the agreed upon list to the ranked list, we noticed that even though half of the fixed subjects coincided, the agreed upon curated list provided an arc of more distinct and precise subjects.

In the second exercise we wanted to understand consistency in applying fixed subjects. For this, three team members with cataloguing experience listened to the same audio and selected up to 10 terms from the Library of Congress Subject Headings to describe them. Comparing the three sets we noticed that each team member had a slightly different

criteria for assigning subjects to the same interview. The three team members' choices coincided only in one subject, pairs of two in four subjects, and pairs of two in five different subjects that could be considered close in meaning. In addition, the three noted that given the range of topics discussed in each interview, they had difficulty between describing the audio in general or in parts.

While these simple exercises should be extended to more participants and analyses of more interviews, they are useful to explain some quantitative results. Even when selecting from controlled vocabularies to describe the same content, catalogers may not coincide in their criteria. That explains the short coverage of the majority of the fixed subjects and consequently the limitations to predict more and diverse labels with higher confidence. Similar to the transcription quality, the diversity and the consistency of the metadata to train the models are important for predicting good labels.

These observations suggest that using LAM metadata to train ML models requires a careful design. Specifically, strategies for assigning metadata should work toward representing the contents of the collection in a more consistent and complete fashion. Based on this study, some suggestions are to further curate the list of controlled vocabularies and provide more structured directions on how to use the hierarchical relations between top scheme and narrower terms. The protocol mentioned by the StoryCorps archivists of applying metadata at the beginning, middle, and end of each interview may render good results if combined with a consistent approach to selecting curated terms. Indeed, more experimenting and testing is needed in order to contribute optimal directions for applying metadata that can be used for training ML models.

B. Interface Design for Lowering Barriers and to Increase Transparency

To lower the technical barriers for LAM to adopt ML we implemented the workflow in IDOLS. The sequence of screens in Fig. 8 below, illustrate how we operationalized share-ability, reproducibility, and transparency as interfaces for the different steps.

In the workflow manager users can select the tasks that they want to run or run them all at once. Enabling granular task management allows verifying the results of each step. The first step presents users with the possibility of selecting files to process and grouping them as needed. At this step, and at any other time in the workflow, users can listen to selected audio files, which is important for comparing with the corresponding transcripts, and for evaluating how the predicted labels represent the audio content qualitatively. For transparency and flexibility, the next task allows users to select executables in the resources in which computations are run, and to configure the IDOLS interfaces script. The latter is useful for creating and improving the forms/interfaces and the directions and explanations about each task. In the following steps, users can run, verify, and compare the transcription results for more than one method (the image shows result from DS), as well as those from the label prediction models. To facilitate the evaluation of label prediction task, the interface shows the results of all the methods per interview and per label including their accuracy assessment values and the fixed subjects assigned by the facilitators for comparison. Lastly, users can provide feedback to the process by submitting a form which can be configured in step 2.

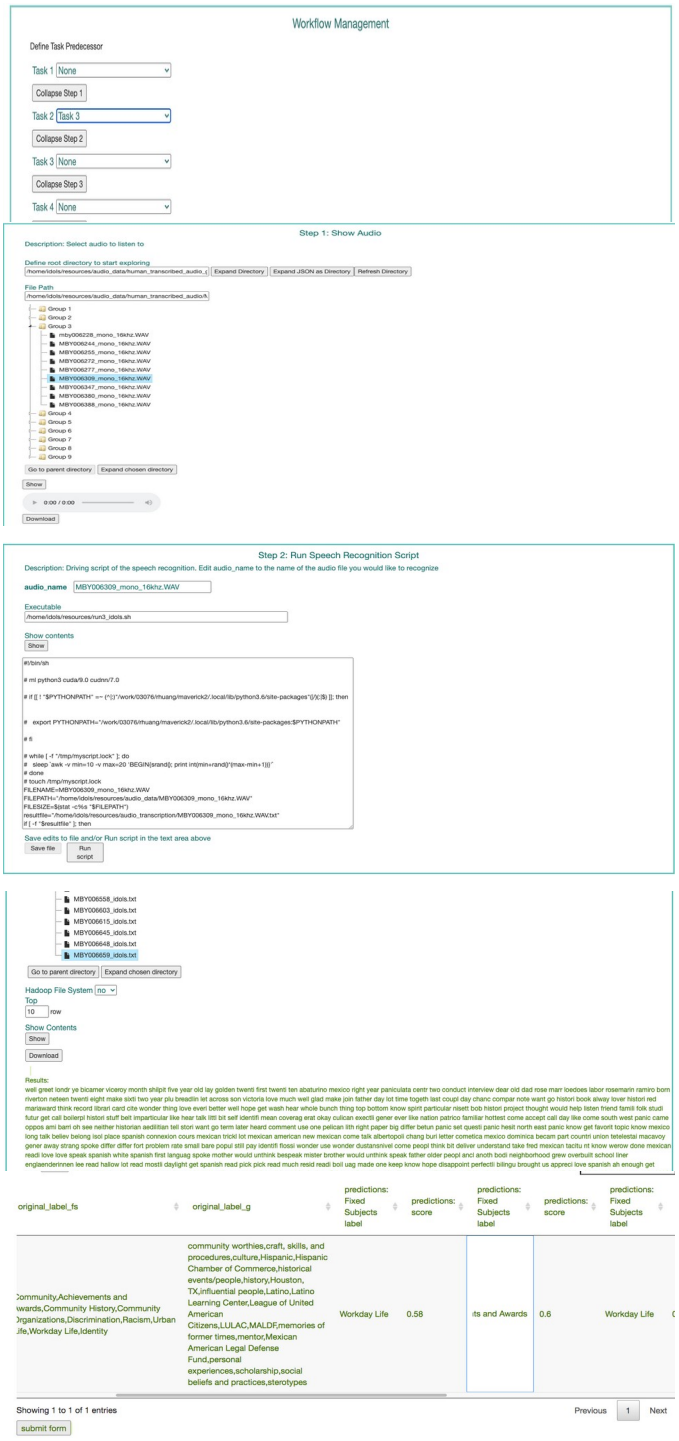

Fig. 8. Sequence of interfaces showing key steps of the ML workflow in IDOLS. (Note: some of the screen captures are cropped to fit and partially shown here.)

The current interfaces are based on the workflow steps and on the feedback provided by the team. The configuration file can be downloaded and shared for reproducibility.

VII. Conclusions

We explored different SRA and ML methods to predict descriptive labels based on transcriptions of speech audio and on subject metadata assigned during archival description. Our work focused on evaluating the accuracy of the labels based on the quality of the automatic transcriptions and of the metadata that was used to build the computational models. A prototype workflow was implemented as a series of interfaces to allow flexible, transparent, and reproducible management of the processing steps. We will continue refining and testing the prototype both from a usability perspective as well as in relation to scalability.

Our interdisciplinary team approached the study with an eye on LAM needs, requirements, and best practices. Because the *Historias* collection and its provenance are both complex and rich, the use case provided the opportunity for addressing real world problems, and challenged the team to combine quantitative and qualitative approaches iteratively. By comparing different ASR methods and using HT as a standard, we learned that it is feasible to use automatic transcripts to generate labels. However, the label's accuracy will depend on the quality of the underlying transcriptions, which can be further improved with more experimentation. Likewise, the quality and coverage of the training metadata is key to achieving good content representation. Using metadata that was not intended to build ML models may not be optimal, but the lessons learned can illuminate future design of training datasets for label prediction.

This study opens new possibilities for LAM that are in sync with professional practices and values. It demonstrates a methodological path for describing the contents of large audio collections that can significantly improve their prompt discoverability. Coupled with a configurable and flexible framework, it can promote cross institutional collaboration and sharing of resources.

The study reveals the advantages of interdisciplinary work. Quantitative methods can help LAM professionals evaluate their metadata practices, and qualitative observations assess the understandability and of the usability of the predicted labels. Indeed, a curatorial perspective is required for ML methods to achieve their higher potential for processing large audio collections.

Acknowledgment

We thank StoryCorps and the Benson Latin American Collection for providing the data and metadata used in this study. We acknowledge the computational resources provided by TACC. The project was funded by UT Austin Good Systems [61], and by the National Science Foundation (Award# 1726816).

References

[1] De Jong, F., Ordelman, R., Scagliola, S., Human Media Interaction, & Faculty of Electrical Engineering, M. (2011). Audio-visual Collections and the User Needs of Scholars in the Humanities: a Case for Co-Development. *Proceedings of the 2nd Conference on Supporting Digital Humanities (SDH 2011)*, (SDH 2011).

[2] Ordelman, R., Heeren, W., Huijbregts, M., de Jong, F., Hiemstra, D., & Ordelman, R. (2009). Towards Affordable Disclosure of Spoken Heritage Archives. Journal of Digital Information, 10(6), NP–NP. Retrieved from http://search.proquest.com/docview/818634813/

[3] Clement, Tanya and Stephen McLaughlin, "Measured Applause: Toward a Cultural Analysis of Audio Collections," Cultural Analytics May 23, 2016. DOI: 10.22148/16.002

[4] Huurnink, B., Hollink, L., Van Den Heuvel, W., & De Rijke, M. (2010). Search behavior of media professionals at an audiovisual archive: A transaction log analysis. *Journal of the American Society for Information Science and Technology*, *61*(6), 1180–1197. https://doi.org/10.1002/asi.21327

[5] Wagstaff, K., & Liu, G. (2018). Automated Classification to Improve the Efficiency of Weeding Library Collections. *Journal Of Academic Librarianship*, *44*(2), 238–247. https://doi.org/10.1016/j.acalib.2018.02.001

[6] Jakeway, E., Algee, L., Allen, L., Ferriter, M., Mears, J., Potter, A., & Zwaard, K. (2020). *Machine Learning + Libraries Summit Event Summary*. Library of Congress. https://blogs.loc.gov/thesignal/2020/02/machine-learning-libraries-summit-event-summary-now-live/

[7] Yige Wang, Ruizhu Huang, and Weijia Xu. 2018. Authentication with User Driven Web Application for Accessing Remote Resources. In Proceedings of the Practice and Experience on Advanced Research Computing (PEARC '18). ACM, New York, NY, USA, Article 2, 7 pages. DOI: https://doi.org/10.1145/3219104.3229290

[8] Weijia Xu, Ruizhu Huang and Yige Wang, "Enabling User Driven Big Data Application on Remote Computing Resources," 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 2018, pp. 4276-4284. doi: 10.1109/BigData.2018.8622006

[9] Hannun, Awni, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger et al. "Deep speech: Scaling up end-to-end speech recognition." *arXiv preprint arXiv:1412.5567* (2014).

[10] Amodei, Dario, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper et al. "Deep speech 2: End-to-end speech recognition in english and mandarin." In *International conference on machine learning*, pp. 173-182. 2016.

[11] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).

[12] StoryCorps. (n.d.). *StoryCorps Historias*. Retrieved April 9, 2020, from https://storycorps.org/discover/historias/

[13] Millington, V., Santiago, M., & Thompson, T. (2019, December 18). *StoryCorps and AI.4.AV Team Interview* (M. Esteva, A. Choate, & T. Clement, Interviewers) [Zoom].

[14] Padilla, T. (2019). *Responsible Operations: Data Science, Machine Learning, and AI in Libraries*. OCLC Research. https://www.oclc.org/research/publications/2019/oclcresearch-responsible-operations-data-science-machine-learning-ai.html

[15] Society of American Archivists. (2020, August). *SAA Core Values Statement and Code of Ethics | Society of American Archivists*. Society of American Archivists. https://www2.archivists.org/statements/saa-core-values-statement-and-code-of-ethics#core_values

[16] "Computational Archival Science." *Computational Archival Science*, https://dcicblog.umd.edu/cas/. Accessed 19 Oct. 2020.

[17] Stanford University Libraries AI Studio "Transcribing the Allen Ginsberg Tapes." https://library.stanford.edu/projects/artificial-intelligence/sul-ai-studio. Accessed 24 Sept. 2020.

[18] Ann Hanlon, Dan Siercks, Marcy Bidney, Cary Costello. LGBTQ+ Audio Archive Mining Project. https://uwm.edu/lib-collections/grant-awarded-for-lgbtq-audio-archive-mining-project/. Accessed 24 Sept. 2020.

[19] Suominen, O. (2019). Annif: DIY automated subject indexing using multiple algorithms. *LIBER Quarterly*, *29*(1), 1–25. DOI: http://doi.org/10.18352/lq.10285

[20] Harish Maringanti. Feasibility of Applying Off-the-Shelf Artificial Intelligence Tools on Digital Library Images Collections – Open Repositories 2021. https://or2020.sun.ac.za/2020/05/29/feasibility-of-applying-off-the-shelf-artificial-intelligence-tools-on-digital-library-images-collections/. Accessed 18 Sept. 2020.
[21] "Collections as Data -Part to Whole.", https://collectionsasdata.github.io/part2whole/. Accessed 24 Sept. 2020.
[22] "The Santa Barbara Statement on Collections as Data." Always Already Computational - Collections as Data, https://collectionsasdata.github.io/statement/. Accessed 23 Sept. 2020.
[23] Reddy, D. Raj. "Speech recognition by machine: A review." Proceedings of the IEEE 64, no. 4 (1976): 501-531.
[24] R.K.Moore, Twenty things we still don't know about speech , Proc.CRIM/ FORWISS Workshop on Progress and Prospects of speech Research an Technology , 1994.
[25] Xinwei Li et.al., Solving large HMM Estimation via Semi-definite programming , IEEE Transactions on Audio, speech and Language processing, Vol.15,No.8, December 2007
[26] Mathias De-Wachter et.al., Template based continuous speech recognition ,IEEE transactions on Audio, speech and Language processing, Vol.15,No.4, May 2007.
[27] Shantanu Chakrabarthy et.al., Robust speech feature extraction by Growth transformation in Reproducing Kernel Hilbert space , IEEE Transactions on Audio, Speech and Language processing, Vol.15,No.6, June 2007.
[28] Shigeru Katagiri et.al., A New hybrid algorithm for speech recognition based on HMM segmentation and learning Vector quantization , IEEE Transactions on Audio, Speech and Language processing Vol.1,No.4,
[29] Tóth, László, and Tamás Grósz. "A comparison of deep neural network training methods for large vocabulary speech recognition." In International Conference on Text, Speech and Dialogue, pp. 36-43. Springer, Berlin, Heidelberg, 2013.
[30] Hinton, Geoffrey, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior et al. "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups." *IEEE Signal processing magazine* 29, no. 6 (2012): 82-97.
[31] Senior, Andrew, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and Brian Kingsbury. "FUNDAMENTAL TECHNOLOGIES IN MODERN SPEECH RECOGNITION." (2012).
[32] Bourlard, Herve, and Nelson Morgan. "Continuous speech recognition by connectionist statistical methods." IEEE Transactions on Neural Networks 4, no. 6 (1993): 893-909.
[33] Renals, Steve, Nelson Morgan, Hervé Bourlard, Michael Cohen, and Horacio Franco. "Connectionist probability estimators in HMM speech recognition." IEEE transactions on speech and audio processing 2, no. 1 (1994): 161-174.
[34] Abdel-Hamid, Ossama, Abdel-rahman Mohamed, Hui Jiang, and Gerald Penn. "Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition." In 2012 IEEE international conference on Acoustics, speech and signal processing (ICASSP), pp. 4277-4280. IEEE, 2012.
[35] Sainath, Tara N., Abdel-rahman Mohamed, Brian Kingsbury, and Bhuvana Ramabhadran. "Deep convolutional neural networks for LVCSR." In 2013 IEEE international conference on acoustics, speech and signal processing, pp. 8614-8618. IEEE, 2013.
[36] Waibel, Alex, Toshiyuki Hanazawa, Geoffrey Hinton, Kiyohiro Shikano, and Kevin J. Lang. "Phoneme recognition using time-delay neural networks." IEEE transactions on acoustics, speech, and signal processing 37, no. 3 (1989): 328-339.
[37] Graves, Alex, Abdel-rahman Mohamed, and Geoffrey Hinton. "Speech recognition with deep recurrent neural networks." In 2013 IEEE international conference on acoustics, speech and signal processing, pp. 6645-6649. IEEE, 2013.
[38] Sak, Hasim, Andrew W. Senior, and Françoise Beaufays. "Long short-term memory recurrent neural network architectures for large scale acoustic modeling." (2014).
[39] Sainath, Tara N., Oriol Vinyals, Andrew Senior, and Haşim Sak. "Convolutional, long short-term memory, fully connected deep neural networks." In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4580-4584. IEEE, 2015.
[40] Mirończuk, Marcin Michał, and Jarosław Protasiewicz. "A recent overview of the state-of-the-art elements of text classification." Expert Systems with Applications 106 (2018): 36-54.
[41] Manning, Christopher D., Hinrich Schütze, and Prabhakar Raghavan. Introduction to information retrieval. Cambridge university press, 2008.
[42] Schenker, A., Bunke, H., Last, M., & Kandel, A. (2005). Graph-theoretic techniques for web content mining. Series in machine perception and artificial intelligence, *Vol.* 62.
[43] Mihalcea, Rada, and Dragomir Radev. Graph-based natural language processing and information retrieval. Cambridge university press, 2011.
[44] Haddoud, Mounia, Aïcha Mokhtari, Thierry Lecroq, and Saïd Abdeddaïm. "Combining supervised term-weighting metrics for SVM text classification with extended term representation." Knowledge and Information Systems 49, no. 3 (2016): 909-931.
[45] Aas, Kjersti, and Line Eikvil. "Text categorisation: A survey." (1999).
[46] Aggarwal, Charu C., and ChengXiang Zhai. "A survey of text classification algorithms." In Mining text data, pp. 163-222. Springer, Boston, MA, 201
[47] Guzella, Thiago S., and Walmir M. Caminhas. "A review of machine learning approaches to spam filtering." Expert Systems with Applications 36, no. 7 (2009): 10206-10222.
[48] Zhang, Zhengyan, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. "ERNIE: Enhanced language representation with informative entities." arXiv preprint arXiv:1905.07129 (2019).
[49] Radford, Alec, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. "Improving language understanding by generative pre-training." (2018): 12.
[50] Golovanov, Sergey, Rauf Kurbanov, Sergey Nikolenko, Kyryl Truskovskyi, Alexander Tselousov, and Thomas Wolf. "Large-scale transfer learning for natural language generation." In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 6053-6058. 2019.
[51] Captioning and Transcription Services | The University of Texas at Austin. https://captioning.lib.utexas.edu/. Accessed 24 Sept. 2020.
[52] DeepSpeech. 2016. Mozilla, 2020. GitHub, https://github.com/mozilla/DeepSpeech.
[53] Abadi, Martín, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin et al. "Tensorflow: A system for large-scale machine learning." In 12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16), pp. 265-283. 2016.
[54] DeepSpeech Model performance https://github.com/mozilla/DeepSpeech/releases/tag/v0.8.2, last accessed Oct. 2020
[55] Speech to Text Web Service, https://cloud.google.com/speech-to-text
[56] Holland, Sarah, et al. "The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards." ArXiv:1805.03677 [Cs], May 2018. arXiv.org, http://arxiv.org/abs/1805.03677.
[57] Manning, Christopher D., Hinrich Schütze, and Prabhakar Raghavan. Introduction to information retrieval. Cambridge university press, 2008.
[58] *The AFS Ethnographic Thesaurus - American Folklore Society*. https://www.afsnet.org/page/AFSET. Accessed 16 Oct. 2020.
[59] Klakow, Dietrich, and Jochen Peters. "Testing the correlation of word error rate and perplexity." Speech Communication 38, no. 1-2 (2002): 19-28.
[60] Dernoncourt, Franck, Trung Bui, and Walter Chang. "A Framework for Speech Recognition Benchmarking." In INTERSPEECH, pp. 169-170. 2018.
[61] University of Texas at Austin, Good Systems, Bridging Barriers Program, https://bridgingbarriers.utexas.edu/projects/building-and-testing-machine-learning-methods-for-metadata-generation-in-audiovisual-collections/ last accessed Oct. 2020