**INITIAL APPLICATION PART 2**

1. **Abstract:**

Content analysis, where annotators assign labels of interest to a set of audio-visual or text documents, is often necessary to understand and characterize language into a dataset that is useful for decision making. Given a large collection of technical, domain-specific texts, it can be very time consuming and expensive for annotators (especially when they need to have expertise in the particular domain) to manually read through the texts to generate appropriate labels with the categories or themes present in the texts, and assign labels to each text. The goal of this project is to develop and evaluate a semi-automated procedure for assisting domain experts in annotating a collection of domain-specific technical documents. Our procedure would enable interaction between automated computational text analysis methods (or models) and the annotators. Using the tool, in an reciprocal manner the annotator's labeling is guided by the output of the computational model, while annotator-provided labels are used to update the computational model. The larger goal is to provide the computational model a small set of high-quality labels that can enable the model to provide reliable automatic predictions for the rest of the documents in the collection, thus potentially making the annotator's task much faster.

To design the most effective interactive annotation tool, we will compare different possibilities for the way a particular set of computational techniques can be deployed. The comparison will be based on various evaluation metrics that will aim to capture the efficiency as well as reliability of the tool, and use a collection of technical documents that have previously been labeled by experts in order to ground the evaluation. The computational techniques include i) a "topic model" that can discover implicit categories, themes, or topics in the data (which can help create groups of documents according to what is being talked about); ii) a "classification model" that can use the words that characterize the particular document in order to predict which label (out of a given set of labels) might be the best fit; and iii) a method that determines which document (or instance of data) should next get labeled (by the annotator) in order to help the classification model become less uncertain of its labeling predictions.

1. **Subject Selection:**
2. **Recruitment:** We will use the online panel platform Prolific to recruit a nationally representative sample of participants.
3. **Eligibility Criteria:** US citizens that are at least 18 years of age, well-versed in the English language (having at least completed high school with English as the primary language of communication) and have a high task completion rate on the platform.
4. **Rationale:** The documents shown in the tool for annotation comes from communities in the US and understanding their content may require experience of living in the US. All the documents are in the English language and contain a fair amount of technical language, requiring at least high school and preferably undergraduate level of studies conducted in English. High rate of previous task completion on the platform ensures a baseline quality for the recruited annotators.
5. **Enrollment Numbers:** In total, we expect to recruit about 500 participants.
6. **Rationale for Enrollment Numbers**: We conducted a power analysis to determine the minimum number of participants needed in various scenarios for the experiment to be well-powered, and give reliable results that can inform our own research as well as future research. On top of the results of our power analysis, we are accounting for unusable annotations and other unexpected noise.
7. **Procedures:**

The research will use a pilot user study to compare different settings for the interactive tool. Each recruited annotator will be assigned an anonymous ID and be randomly put into one of five groups, based on five different annotation scenarios the study seeks to compare. Each group will be shown instructions on how the interface can be used and the layout specific to their specific scenario. In each scenario, participants will be shown a display of their progress, a preview of passages (or documents) that they have to annotate, and the labels they have created at any given stage. Participants will be able to create new labels as well as use the labels they have already created as they annotate documents. Further, the tool allows participants to rename or delete a label they created earlier as they annotate and the changes will automatically be propagated and be reflected in their annotation task.   
  
The following table summarized the five scenarios (each corresponds to one of the groups a participant will be assigned to):

|  |  |  |  |
| --- | --- | --- | --- |
| Scenario | Is topic modeling used? | Is active learning used? | Description |
| 1 |  |  | * Classic manual content analysis and annotation with no assistance from computational tools of topic modeling or active learning * All documents will be shown as a list, participants will be asked to create a label and label each document * The next document to be labeled will be highlighted by a red box in random order * Participants will be able to go back and forth as they label the documents, create new labels, use existing ones - this basically functionality of the tool will remain consistent across scenarios |
| 2 |  |  | * Documents shown will be grouped by topics or implicit categories discovered by a topic model * The participants can see the top characteristics words or terms associated with that topic * The tool will suggest the next document to label (via highlighting with a red box) at random, but each document will be a part of a group informed by the topic model as well as have the top characteristic word for that discovered category or topic shown |
| 3 |  |  | * Documents will be shown as a list, but after a minimum number of documents have been annotated and a minimum number of distinct categories have been created by the user, the next document that the user should label will be selected using active learning * The next document for annotation will be highlighted (with a red box) based on uncertainty score of the classifier - the document that the model is most uncertain about in its label prediction |
| 4 |  |  | * On top of documents shown in groups based on topics given by topic model as well as the top characteristic words for the topic in scenario 2, the next document highlighted (using a red box) for annotation is selected using active learning (as done in scenario 3) instead of random selection |
| 5 |  |  | * Participants will be shown the top characteristic words as well as documents for each topic or discovered category in the collection. * Instead of annotating documents, the participants will create a label for that topic by reading the documents grouped under that topic and the top characteric words. This label will be one the participants feel comfortable applying to all documents shown in that particular group, and the annotation of those particular group of documents will happen automatically once a label has been created for that topic. * The participant will then be asked to move to a different topic and group of documents randomly. * A time limit will of no more than 15 minutes will be set for coming up with a label for a particular group, but participants can revise and modify labels at the end of annotation process once they have had a chance to look at all the documents. |

The evaluation will only use the labels created by the participant, used to annotate documents or passages shown to them (coming from a collection of technical planning documents). At any stage of annotation, certain documents would have been annotated or labeled and others would have not. The annotated documents will be used by the computational model (machine learning classifier) to assign the labels created by that point to all the other unlabeled documents in the collection, and we will compare the groups of documents formed by these ‘predicted’ labels with the actual ground-truth labels. The best setting should achieve high agreement with ground truth labels with least amount of documents annotated by human users using the interactive tool (potentially increased *efficiency*). Further, the *reliability* of the tool as a way to assist document annotation can be established by looking at the inter-rater agreement, measured by similarity of the resulting groups of documents as per the annotations by participants in each of the various groups. The tool will only record the labels created and assigned to the documents, and calculate the comparison of predicted labels for the whole collection with ground truth at every new annotation (which will then give us the trajectory of this alignment rate with number of documents annotated). The final set of annotations from every participant will be used for computing inter-rater agreement, and the average agreement can be used to compare the five different scenarios. **No de-anonymizing data of the participant will be collected, only their annotations.**

Participants will be paid through the online panel platform used for the recruiting itself (Prolific). Each annotator will be allotted one hour of total time for giving consent, reading the instructions and examples to understand the task, complete their specific annotations, and fill out a quick post-participation questionnaire. Each participant will be paid USD 20 for one hour of their time through Prolific. Any suspected fraud or abuse will result in forfeiture of compensation, and that includes not actually completing the task (more than 5 instances left unlabeled).

The post-participation questionnaire is included in the supporting documents and aims to assess the user experience in terms of how demanding the task was, how confident they felt about their annotations, how helpful they would find the interface for its goal of helping annotate documents, and their feedback on the tool.

1. **Risks:**

There are no known risks to the participants.

1. **Benefits:**

There are no direct benefits to the participants. Their participation will help determine the way an interactive tool can be used by humans seeking to annotate a large set of documents with no given label set (content analysis) - a crucial and widespread method of understanding the content of a collection of texts.

1. **Confidentiality:**

No identifying information will be collected or retained. Users may be asked to create (or be assigned) an anonymous ID in order to save their work in the time they are using the tool. Any post-participation surveys will only be used to understand their experience during the task and not ask for any identifying information. Unless users disclose any personally identifiable information in their labels (they will be reminded not to), such information would not be obtained, and we will review all annotation data to ensure that any such information is removed and completely anonymized before any public release of the labels created during annotation. Labels prior to anonymized public release and all other data will be stored securely in a password-protected account, and only the principal investigator and co-investigators will have access to it to minimize any loss of confidentiality.

1. **Consent Process:**

We are collecting anonymous, non-identifiable annotations and to this end, we request a waiver of informed consent requirement on the following grounds -

1) The research involves no more than minimal risk to the subjects

The research will involve annotation of technical documents using an interactive tool, and answering a few survey questions. All of this will be online. This will not create any risks that are greater in and of themselves than those ordinarily encountered in daily life.

(2) The waiver or alteration will not adversely affect the rights and welfare of the subjects

The participation in the study will be completely anonymous, and no identifying information for any participant will be recorded.

(3) The research could not practicably be carried out without the waiver or alteration

The annotation task will be conducted online and therefore hardcopy documentation of written consent is not feasible.

(4) Whenever appropriate, the subjects will be provided with additional pertinent information after participation.

Participants can obtain additional information after participation by contacting the principal investigator (contact details are included in the recruitment material as well as the consent document provided as supporting documents).

Implied consent will be obtained from participants by them reviewing the text included in the consent document (supporting materials), and clicking a button to continue on with actually beginning the task. Participants may download or print a copy of the consent document for their records.

1. **Conflict of Interest:**

No conflict of interest. One of the funding sources of the project is Adobe. However, Adobe will not be involved in any capacity in this particular user study and will not have privileged access to the data.

1. **HIPAA Compliance:**

Not applicable.

1. **Research Outside of the United States:**

Not applicable.

1. **Research Involving Prisoners:**

Not applicable.

1. **SUPPORTING DOCUMENTS**

* **completed Initial Application Part 1 (On-Line Document)**
* **post-participation survey questionnaire document**
* **consent document**
* **recruitment document**