Speeding Document Annotation with Topic Models

Forough Poursabzi-Sangdeh and Jordan Boyd-Graber
Computer Science
University of Colorado Boulder
{forough.poursabzisangdeh, Jordan.Boyd.Graber}@colorado.edu

Abstract

Document classification and topic models are useful tools for managing and understanding large corpora. Topic models are used to uncover underlying semantic and structure of document collections. Categorizing large collection of documents requires hand-labeled training data, which is time consuming and needs human expertise. We believe engaging user in the process of document labeling helps reduce annotation time and address user needs. We present an interactive tool for document labeling. We use topic models to help users in this procedure. Our preliminary results show that users can more effectively and efficiently apply labels to documents using topic model information.

1 Introduction

Many fields depend on texts labeled by human experts; computational linguistics uses such annotation to determine word senses and sentiment (Kelly and Stone, 1975; Kim and Hovy, 2004); social science uses "coding" to scale up and systemetize content analysis (Budge, 2001; Klingemann et al., 2006). In general text classification is a standard tool for managing large document collections.

However, these labeled data have to come from somewhere. The process for creating a broadly applicable, consistent, and generalizable label set and then applying them to the dataset is long and difficult, requiring expensive annotators to examine large swaths of the data.

We present a user interactive tool for document labeling that uses topic models to help users assign appropriate labels to documents (Section 2). In Section 3, we describe our user interface and experiments on Congressional Bills data set. We also explain an evaluation metric to assess the quality of assigned document labels. In preliminary results, we show that annotators can more quickly label a document collection given a topic modeling overview. While engaging user in the process of content-analysis has been studied before(as we discuss in Section 4), in Section 4 we describe how our new framework allows for more flexibility and interactivity. Finally, in Section 5, we discuss the limitation of our framework and how we plan to extend it in future.

2 Interactive Document Labeling

We propose an alternative framework for assigning labels to documents. We use topic models to give an overview of the document contents to the user. Users can create a label set incrementally, see the content of documents, assign labels to documents, and classify documents. They can go back and forth in these steps and edit label set or document labels and re-classify.

Having labeled documents is necessary for automatic text classification. With a large collection of unstructured documents, labeling can be excruciating since it is essential to label enough documents in different labels to obtain acceptable accuracy. Topic models are a solution to reduce this effort since they provide some information about the underlying theme of corpus. Given a fixed number of topics, topic models

Proceedings of NAACL-HLT 2015 Student Research Workshop (SRW), pages 126–132,
Denver, Colorado, June 1, 2015.

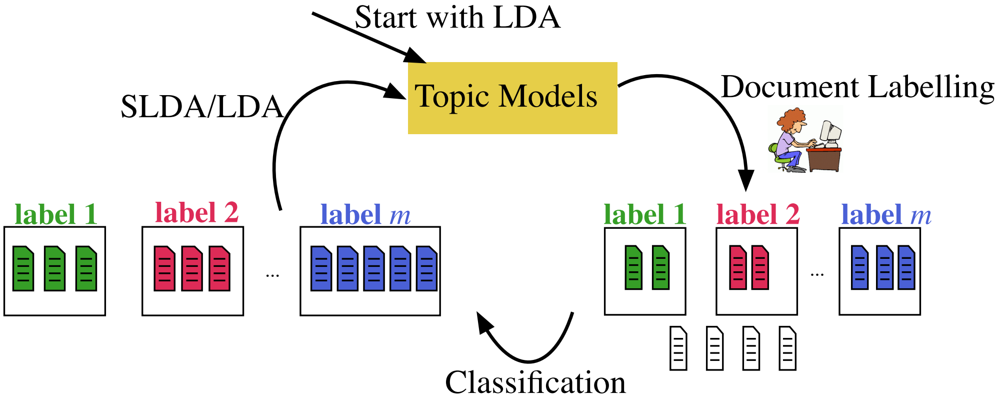


Figure 1: Interactive document labeling: Start with LDA topic modeling, show users relevant documents for each topic, get user labels, classify documents, and use SLDA to generate topics[1]. Repeat this until the user is satisfied with labels.

output (i) a set of words for each topic (*Topic words*) and (ii) a distribution over topics for each document (*Document's Topic Distribution*).

Topic words can be used to reveal the content of a topic and thus content of documents with a high probability of that topic. Therefore, assuming the number of topics is chosen carefully, top documents for each topic are similar in content and can be labeled appropriately.

Thus, rather than showing an unstructured collection of documents to the user, providing the topic words and highly relevant documents to that topic helps them in the process of document labeling, both in the step of choosing appropriate label names and choosing appropriate document to assign a label to. Another way to think about this is that if the topics are perfect (they are not too general or too detailed), all labels associated with the topic's high relevant documents can be viewed as subjects explaining the topic. Table 1 provides an example of how topic models can help a user craft document labels.

Having a set of user labeled documents, classification algorithms can be used to predict the label of unseen documents. Next, classification results are shown. Users can change document labels. They can also edit/delete label set and re-run the classifier. The explained procedure can be repeated iteratively until satisfaction is achieved with existing (*document,label*) pairs. Figure 1 shows the explained procedure.

3 Experiments with Interactive Labeling Interface

Data: In our experiments, we need a labeled corpus to be able to assess the quality of user-generated labels. We chose US Congressional Bills corpus (Adler and Wilkerson, 2006). GovTrack provides bill texts along with the discussed congressional issues as labels. Example of labels are "education", "agriculture", "health", and "defense". There are total of 19 unique labels. We use the 112th congress, which has 12274 documents. We remove bills with no assigned gold label or that are short. We end with 6528 documents.

Topic Modeling: To generate topics, we use Mallet (McCallum, 2002) to apply LDA on the data. A set of extra stop words are generated based on TF-IDF scores to avoid displaying non-informative words to the user.

Features and Classification: A crucial step for text classification is to extract useful features to represent documents. Some common features for text classification are n-grams, which makes the dimensionality very high and classification slower. Since response time is very important in user interactive systems, instead of n-grams, we

[1] Currently, we are not using SLDA. We just use the original topics generated by LDA. The idea behind SLDA is explained in Section 5.

Topic	Words	Document Title	Document Labels
16	dod, sbir, afghanistan, phase, sttr, missile, combat, capabilities, command, elements	HR 4243 IH 112th CONGRESS 2d Session H. R. 4243 To strengthen the North Atlantic Treaty Organization.	military
19	historic, conveyance, dated, monument, depicted, generally, boundary, creek, preservation, recreation	HR 4334 IH 112th CONGRESS 2d Session H. R. 4334 To establish a monument in Dona Ana County, New Mexico, and for other purposes.	wildlife
		S 617 IS 112th CONGRESS 1st Session S. 617 To require the Secretary of the Interior to convey certain Federal land to Elko County, Nevada, and to take land into trust for the Te-moak Tribe of Western Shoshone Indians of Nevada, and for other purposes.	nature

Table 1: An example of topic words and the labels user has assigned to top documents for that topic.

use topic probabilities as features, which reduces the dimensionality and classification time significantly. User can choose 10, 15, 25, or 50 topics. We want to show the label probabilities generated by classifier to users. We use Liblinear (Fan et al., 2008) to run L2 regularized logistic regression for classifying documents and generating label probabilities.

Interface: We start with the web-based interface of Hu et al. (2014) for interactive topic modeling. The existing interface starts with asking user information, corpus name, and number of topics they want to explore. Then it displays topic words and the most relevant documents for each topic. Also, the user can see the content of documents. Users can create new labels and/or edit/delete an existing label.

When seeing a document, user has 3 options:

1. Create a new label and assign that label to the document.
2. Choose an existing label for the document.
3. Skip the document.

At any point, the user can run the classifier. After classification is finished, the predicted labels along with the certainty is shown for each document. User can edit/delete document labels and re-run classifier as many times as they desire. We Refer to this task as *Topic Guided Annotation*(TGA).

Figure 2 shows a screenshot of the interface when choosing a label for a document.

3.1 Evaluation

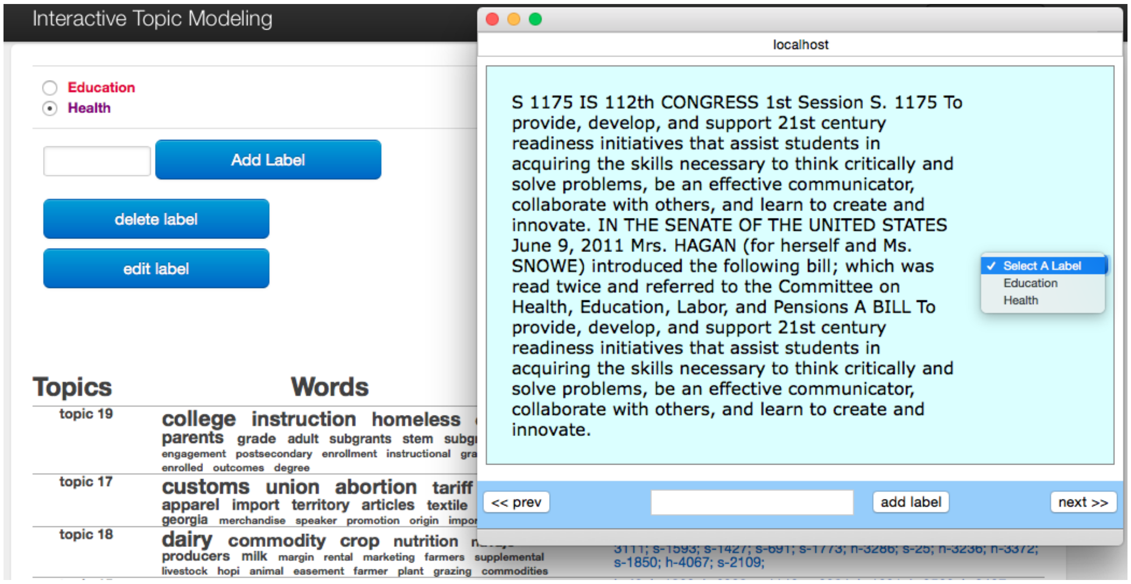
We introduce an interactive framework for document labeling using topic models. In this section, we evaluate our system.

Our goal is to measure whether showing users a topic modeling overview of the corpus helps them apply labels to documents more effectively and efficiently. Thus, we compare user-generated labels (considering labels assigned by user and classifier altogether) with gold labels of US Congressional Bills provided by GovTrack. Since user labels can be more specific than gold labels, we want each user label to be “pure” in gold labels. Thus, we use the purity score (Zhao and Karypis, 2001) to measure how many gold labels are associated with each user label. Purity score is

$$\text{purity}(\mathscr{U},\mathscr{G}) = \frac{1}{N}\sum_k \max_j |U_k \cap G_j|, \qquad (1)$$

where $\mathscr{U} = \{U_1, U_2, ..., U_K\}$ is the user clustering of documents, $\mathscr{G} = \{G_1, G_2, ..., G_J\}$ is gold clustering of documents, and N is the total number of documents. Moreover, we interpret U_k and G_j as the set of documents in user cluster U_K or gold cluster G_j. Figure 3 shows an example of purity calculation for a clustering, given gold labels.

Purity is an external metric for cluster evaluation. A very bad labeling has a purity score close to 0 and a perfect labeling has purity score of 1.



S 1175 IS 112th CONGRESS 1st Session S. 1175 To provide, develop, and support 21st century readiness initiatives that assist students in acquiring the skills necessary to think critically and solve problems, be an effective communicator, collaborate with others, and learn to create and innovate. IN THE SENATE OF THE UNITED STATES June 9, 2011 Mrs. HAGAN (for herself and Ms. SNOWE) introduced the following bill; which was read twice and referred to the Committee on Health, Education, Labor, and Pensions A BILL To provide, develop, and support 21st century readiness initiatives that assist students in acquiring the skills necessary to think critically and solve problems, be an effective communicator, collaborate with others, and learn to create and innovate.

Topics	Words
topic 19	college instruction homeless parents grade adult subgrants stem subg engagement postsecondary enrollment instructional gra enrolled outcomes degree
topic 17	customs union abortion tariff apparel import territory articles textile georgia merchandise speaker promotion origin impor
topic 18	dairy commodity crop nutrition producers milk margin rental marketing farmers supplemental livestock hopi animal easement farmer plant grazing commodities

3111; s-1593; s-1427; s-691; s-1773; h-3286; s-25; h-3236; h-3372; s-1850; h-4067; s-2109;

Figure 2: A screenshot of interactive document labeling interface. The user sees topic words and the most relevant documents for each topic. The user has created two labels: “Education” and “Health” and sees the content of a documents. The user can create a new label and assign the new label to the document, or choose one of the two existing labels to assign to the document, or skip the document and view the previous or next document.

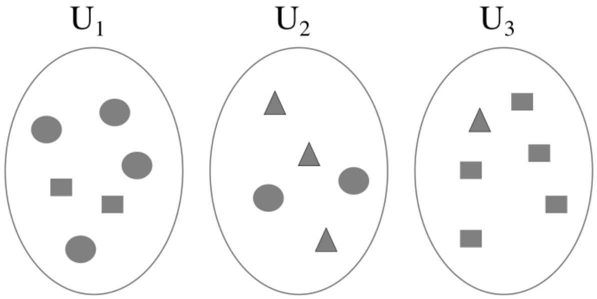


Figure 3: An example of computing purity: Clusters correspond to user labels and different shapes correspond to different associated gold labels. Majority gold label numbers for three clusters are $4(U_1)$, $3(U_2)$, and $5(U_3)$. Purity is $\frac{1}{17} \times (4+3+5) \approx 0.71$.

The higher this score, the higher the quality of user labels.

To evaluate TGA, We did a study on two different users. For User 1, we chose 15 topics and for User 2, we chose 25 topics. They were asked to stop labeling whenever they were satisfied with the predicted document labels.

We compare the user study results with a baseline. Our baseline ignores topic modeling information for choosing documents to labels. It considers the scenario when users are given a large document collection and are asked to categorize the documents without any other information. Thus, we show randomly chosen documents to users and want them to apply label to them. All users can go back and edit or delete document labels, or refuse to label a document if they find it confusing. After each single labeling, we use the same features and classifier that we used for user study with topic models to classify documents. Then we calculate purity for user labels with respect to gold labels. Figure 4 shows the purity score over different number of labeled documents for User 1, User 2, and baseline.

User 1 did the labeling in 6 rounds, whereas User 2 did total of 7 rounds. User 1 ended with 116 labeled documents and user 2 had 42 labeled documents in the end.

User 2 starts with a label set of size 9 and labels 11 documents. Two documents are labeled as “wildlife”, other two are labeled as “tax”, and all other documents have unique labels. This means that even if there are very few instance per label, baseline is outperformed. This is an evidence of

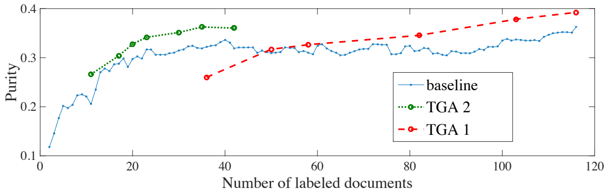


Figure 4: Purity score over number of labeled documents. TGA 1 and TGA 2 refer to results for User 1 and User 2.

User 1	Baseline
36	12
50	52
58	60
82	109
103	> 116
116	> 116

(a)

User 2	Baseline
11	12
17	18
20	38
23	40
30	112
35	116
42	115

(b)

Table 2: The number of required labeled documents for baseline to get the same purity score as (a) User 1 (b) User 2, in each round

choosing informative documents to assign labels with the help of topic models. On the other hand, User 1 starts with a label set of size 7 and labels 36 documents and is outperformed by baseline significantly. One reason for this is that assigning too many documents relevant to a topic, with the same label doesn't provide any new information to the classifier and thus the user could get the same purity score with a lower number of labeled documents, which would lead to outperforming baseline. User 1 outperforms the baseline in the second (8 labels and 50 labeled documents) and third round (9 labels and 58 labeled documents) slightly. In the fourth round, user creates more labels. With total of 13 labels and 82 labeled documents, the gap between user's purity score and baseline gets larger. Both users outperform baseline in the final round.

To see how topic models help speed up labeling process, we compare the number of user labeled documents with the approximate number of required labeled documents to get the same purity score in baseline. Table 2 shows the results for User 1 and User 2.

User 1 starts with man labeled documents and baseline can achieve the same performance with one third of the labeled documents. As the user keeps labeling more documents, the performance improves and baseline needs more labeled documents to get the same level of purity. For User 2, baseline on average needs over two times as many labeled documents to achieve the same purity score as user labels. These tables indicate that topic models help users choose documents to assign labels to and achieve an acceptable performance with fewer labeled documents.

4 Related Work

Topic Models such as Latent Dirichlet Allocation (Blei et al., 2003, LDA) are unsupervised learning algorithms and are a useful tool for understanding the content of large collection of documents. The topics found by these models are the set of words that are observed together in many documents and they introduce correlation among words. Top words in each topic explain the semantics of that topic. Moreover, each document is considered a mixture of topics. Top topics for each document explain the semantics of that document.

When all documents are assigned a label, supervised topic models can be used. SLDA (Mcauliffe and Blei, 2008) is a supervised topic model that generates topics that give an overview of both document contents and assigned labels. Perotte et al. (2011) extend SLDA and introduce HSLDA, which is a model for large-scale multiply-labeled documents and takes advantage of hierarchical structure in label space. HSLDA is used for label prediction. In general, supervised topic models help users understand labeled document collections.

Text classification predicts a label for documents and help manage document collections. There are known classifiers as well as feature extraction methods for this task. However, providing an initial set of labeled documents for both text classification and supervised topic models still requires lots of time and human effort.

Active learning (Settles, 2010), reduces the amount of required labeled data by having a learner which actively queries the label for specific documents and collects a labeled training set. In a user interactive system, the active learner queries document labels from users (Settles, 2010). In other words, the learner suggests some documents to the user and wants the user to assign a label to those. Settles (2011) discusses that having interactive users in annotation process along with active learning, reduces the amount of annotation time while still achieving acceptable performance. In more detail, they presents an interactive learning framework to get user annotations and produce accurate classifiers in less time. The shortcoming of active learning is that they don't provide any overview information of corpus, like topic model approaches do.

Nevertheless, new methods in both analysis and evaluation are needed. Classification algorithms restrict document labels to a predefined label set. Grimmer and Stewart (2013) show that to be able to use the output of automatic text analysis in political science, we need careful validation methods. There has been some work done on bringing user in this task for refining and evaluating existing methods. Hu et al. (2014) show that topic models are not perfect from the user view and introduce a framework to interactively get user feedback and refine topic models. Chuang et al. (2013) present an interactive visualization for exploring documents by topic models to address user needs.

We bring these tools together to speed up annotation process. We believe having users engaged in content analysis, not only reduces the amount of annotation time, but also helps to achieve user satisfaction. We propose an iterative and user interactive procedure for document annotation. We use topic models to provide some high-level information about the corpus and guid users in this task. We show top words and documents for each topic to the user and have them start labeling documents. Users can create/edit/delete labels. Then users can run a classifier to predict the labels for the unlabeled documents. They can change document labels and re-classify documents iteratively, until satisfaction is achieved.

5 Future Work

There are some obvious directions that will expand this ongoing research. First, we are planning to use active learning to better aid classification. We expect that active learning will reduce the number of required labeled documents while still getting a high purity score and user satisfaction.

Second, we will use supervised topic models (Mcauliffe and Blei, 2008, SLDA) instead of LDA after the first round to update topics based on document labels. SLDA uses labeled documents to find topics that explain both document content and their associated labels. We believe using SLDA instead of LDA after the first round will give users more information about the overview of documents and help them further for applying labels to documents.

Third, we want to allow the user to refine and correct labels further. Our existing interface allows the user to delete a label or edit a label. We believe it is also important for users to merge labels if they think the labels are too specific. In addition, we believe a crucially important step is to generate the label set. Giving the user some information about the range of documents can help them generate a better label set. One other option is to suggest labels to users based on topic models (Lau et al., 2010).

Fourth, we will explore other corpora such as European Parliament corpus (Koehn, 2005). To our knowledge, there are no true labels for Europarl corpus and using our interactive tool can help users find the categorized information they need.

Finally, for evaluating our method, in addition to using the correct labeling and purity score, we will conduct a user experiment with more users involved. Since the task of labeling congress data set requires some political knowledge, we will choose annotators who have some political science background.

Acknowledgments

We thank the anonymous reviewers for their insightful comments. We thank Dr. Niklas Elmqvist for his advice for revising the user interface. We also thank Alvin Grissom II for helping us in the user study. This work was supported by NSF Grant NCSE-1422492. Any opinions, findings, results, or recommendations expressed here are of the authors and do not necessarily reflect the view of the sponsor.

References

E Scott Adler and John Wilkerson. 2006. Congressional bills project. *NSF*, 880066:00880061.

David M. Blei, Andrew Ng, and Michael Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3.

Ian Budge. 2001. *Mapping policy preferences: estimates for parties, electors, and governments, 1945-1998*, volume 1. Oxford University Press.

Jason Chuang, Yuening Hu, Ashley Jin, John D Wilkerson, Daniel A McFarland, Christopher D Manning, and Jeffrey Heer. 2013. Document exploration with topic modeling: Designing interactive visualizations to support effective analysis workflows. In *NIPS Workshop on Topic Models: Computation, Application, and Evaluation.*

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874.

Justin Grimmer and Brandon M Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, page mps028.

Yuening Hu, Jordan Boyd-Graber, Brianna Satinoff, and Alison Smith. 2014. Interactive topic modeling. *Machine learning*, 95(3):423–469.

Edward F Kelly and Philip J Stone. 1975. *Computer recognition of English word senses*, volume 13. North-Holland.

Soo-Min Kim and Eduard Hovy. 2004. Determining the sentiment of opinions. In *Proceedings of the 20th international conference on Computational Linguistics*, page 1367. Association for Computational Linguistics.

Hans-Dieter Klingemann, Andrea Volkens, Judith Bara, Ian Budge, Michael D McDonald, et al. 2006. *Mapping policy preferences II: estimates for parties, electors, and governments in Eastern Europe, European Union, and OECD 1990-2003.* Oxford University Press Oxford.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

Jey Han Lau, David Newman, Sarvnaz Karimi, and Timothy Baldwin. 2010. Best topic word selection for topic labelling. In *Coling 2010: Posters*, pages 605–613, Beijing, China, August.

Jon D Mcauliffe and David M Blei. 2008. Supervised topic models. In *Advances in neural information processing systems*, pages 121–128.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. http://www.cs.umass.edu/ mccallum/mallet.

Adler J. Perotte, Frank Wood, Noemie Elhadad, and Nicholas Bartlett. 2011. Hierarchically supervised latent Dirichlet allocation. In *Advances in Neural Information Processing Systems 24*, pages 2609–2617.

Burr Settles. 2010. Active learning literature survey. *University of Wisconsin, Madison*, 52(55-66):11.

Burr Settles. 2011. Closing the loop: Fast, interactive semi-supervised annotation with queries on features and instances. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1467–1478. Association for Computational Linguistics.

Ying Zhao and George Karypis. 2001. Criterion functions for document clustering: Experiments and analysis. Technical report, Citeseer.