Exploring *Jane Addams Papers Project* Documents Through Topic Modeling and Multilabel Classification

By

Olivia Church, Bachelor of Science in Mathematics

A thesis submitted to the Graduate Committee of

Ramapo College of New Jersey in partial fulfillment

of the requirements for the degree of

Master of Science in Applied Mathematics

Spring, 2025

Committee Members:

Dr. Amanda Beecher, Advisor

Dr. Cathy Moran Hajo, Reader

Dr. Debbie Yuster, Reader

COPYRIGHT

Dedication

To my family, whose support, love, and guidance have shaped me into the person I am today

Acknowledgments

I would like to thank Dr. Beecher for being such a supportive and helpful advisor throughout this project. I would not have been able to complete this thesis without your unwavering patience and enthusiasm. Your passion for mathematics is inspiring, and I feel lucky to have been your student.

Thank you to Dr. Hajo for serving as a member of my thesis committee and for allowing me to use data from the *Jane Addams Papers Project*. Your help throughout this project was invaluable. I am grateful to have had the opportunity to learn more about the *Jane Addams Papers Project* and the work that you do.

I would also like to thank Dr. Yuster for being a member of my thesis committee. I appreciate your guidance and the time you dedicated to offering aid and advice.

Lastly, thank you to all of the faculty who have contributed to my education at Ramapo College. I will be forever grateful for my time spent here and all that I have learned.

Table of Contents

List of Tables

List of Figures

Abstract

The *Jane Addams Papers Project* at Ramapo College of New Jersey compiles documents relating to Jane Addams. An American activist and social worker, Addams was an influential member of many political and social movements throughout the nineteenth and twentieth centuries, advocating for women's suffrage, child labor reform, and peace, among other matters. The Digital Edition of the *Jane Addams Papers Project* contains digital versions of the documents, as well as a variety of other features, such as tags that categorize the documents based on their content. To explore new ways of analyzing and organizing documents from the Digital Edition, two machine learning techniques were implemented: topic modeling and multilabel classification. In addition to extracting insights from the documents and developing an automated method of assigning tags, a central aim of this research was to investigate how topic modeling and multilabel classification can be bridged to enrich analyses of texts.

Using a subset of documents from the Digital Edition, speeches and articles written by Jane Addams, latent Dirichlet allocation (LDA) topic modeling identified central topics, or themes, including international affairs and conflicts, child labor, and women's suffrage. A variety of multilabel classification models were utilized to predict tags. The problem transformation algorithm Binary Relevance used in conjunction with a Multinomial Naive Bayes classifier had the best performance, though a higher accuracy would have been more desirable. To link the topic modeling and multilabel classification results, each document and tag was assigned to a specific topic. A connection between the topics and predicted tags of documents was evident, with the multilabel classifier often predicting tags related to the topic of their corresponding document. Therefore, when used together, topic modeling and multilabel classification may

complement each other, potentially contributing to a greater understanding of the subject matter of texts.

Introduction

The *Jane Addams Papers Project* at Ramapo College of New Jersey compiles documents relating to Jane Addams (https://janeaddams.ramapo.edu/). Born in 1860, Addams was an American reformer, activist, and social worker. After co-founding the settlement home Hull-House in Chicago in 1889, which provided services to the working-class community, Addams became involved in many reform movements of the late nineteenth and early twentieth centuries, such as the ones promoting child labor reform and better working conditions ("About Jane Addams," n.d.). Addams was also a prominent member of the women's suffrage movement, serving as a vice president of the National American Woman Suffrage Association. When World War I broke out, Addams advocated for peace, eventually becoming the president of the Women's International League for Peace and Freedom. Addams led a life dedicated to social service, making lasting contributions to the movements of which she was a part until her death in 1935.

As a result of Addams' influential life, there are numerous documents connected to her, including newspaper articles about her, speeches she delivered, and correspondence exchanged between her and others. The goal of the *Project*, which was founded in 1975 at the University of Illinois at Chicago by Mary Lynn Bryan, is to compile these documents into the six-volume work, *Selected Papers of Jane Addams* ("About the Project," n.d.). Bryan completed the first three of these volumes, which contain documents from the years 1860 to 1900. In 2015, Dr. Cathy Moran Hajo brought the *Project* to Ramapo College of New Jersey with the aim of completing the final three volumes in the *Selected Papers of Jane Addams*, spanning the years 1901 to 1935. An additional aim of the *Project* at Ramapo College is to digitize the documents

making up the final three volumes, which has led to the creation of the *Jane Addams Papers Project* Digital Edition, where the documents are "freely available and searchable" online ("About the Project," n.d.). The Digital Edition also houses information about the "people, places, events, and organizations" that connect to Addams and appear in the documents ("About the Project," n.d.).

Machine-based methods can facilitate an analysis of the vast collection of texts contained in the *Jane Addams Papers Project* Digital Edition. The advancement of technology has enabled machines to better understand human language. The field of computer science known as natural language processing (NLP) "has become an indispensable part of modern computing as it empowers computers with intelligence in interpreting textual data" (Kholwal, 2023, p. 416). Two widely used NLP techniques are topic modeling and text classification. Topic modeling is an unsupervised machine learning method that discovers topics, or themes, within a collection of texts (Li et al., 2024). Useful for extracting patterns and providing a sense of the main ideas in texts, topic modeling is "an important strategy enabling conceptual comprehension of text content" (Muthusami et al., 2024, p. 1). Text classification is a similarly important technique, involving the use of supervised machine learning algorithms to assign texts with predefined labels (Voskergian et al., 2024). When more than one label can be assigned to texts, this is known as multilabel classification and "is one of the key problems of modern day Machine Learning" (Tandon & Chatterjee, 2022, p. 4425).

Topic modeling and text classification are valuable tools for managing and analyzing textual data that have been employed in various domains. The central aim of this research is to implement these two techniques on texts from the Digital Edition of the *Jane Addams Papers Project*. The Digital Edition provides online access to not only *Project* documents, but to

information relating to their contents, such as descriptions of the documents and associated labels, or tags. Topic modeling can reveal themes in the documents, uncovering patterns that may not have been previously recognized or that would have been difficult to identify manually as a result of the vastness of the Digital Edition's collection. The topics that result from the model can supplement the current pieces of information that summarize the contents of the documents, like the descriptions and tags, serving as an additional pathway to examine and analyze the texts. This would contribute to a rich, multifaceted environment for exploring the *Jane Addams Papers Project* documents.

While topic modeling can provide insight about entire collections of documents, a text classification system could be used to analyze the contents of individual documents, aiding in the process of assigning labels to them. Since documents in the Digital Edition usually have more than one tag assigned to them, this classification system would be a multilabel one. Automating the categorization of documents would streamline the process, which is currently done by hand. Instead of replacing the process of manually tagging the documents completely, a multilabel classification system could also act as a tool for *Project* staff members, suggesting possible labels for documents that could be taken into consideration. In a similar manner as the topic modeling, the multilabel text classification can serve as a supplemental, machine-based method for analyzing the documents, introducing new ways of examining and organizing them.

Both topic modeling and text classification can be used to process textual data in meaningful ways, with a variety of practical uses. Topic modeling "is widely employed in various fields, including natural language processing, information retrieval, text mining and computational linguistics" (Detthamrong et al., 2024, p. 449). Similarly, text classification "is a fundamental issue in natural language processing, including information retrieval, information

extraction, and text mining" (Elghazel et al., 2016, p. 1). The overlap among the applications of topic modeling and text classification, such as information retrieval and text mining, indicates that the two methods can produce similar kinds of insights about texts. However, existing research does not showcase many ways in which topic modeling and text classification can be used together to explore texts. Therefore, another aim of this study is to investigate not only the kinds of meaning that topic modeling and multilabel text classification can produce separately, but also the kinds of meaning they can produce together when used alongside each other to analyze a collection of documents.

The process of implementing topic modeling and multilabel text classification on the *Jane Addams Papers Project* texts is documented and broken down into different sections in this paper. First, the "Background" section provides an overview of related research. The "Methodology" section outlines the steps taken to implement topic modeling and multilabel text classification and compare the two techniques, while the "Analysis and Discussion" section examines and interprets the results. Finally, the "Conclusions" section provides further discussion of the results and their significance.

Background

Multilabel text classification and topic modeling have been used in various studies to analyze collections of texts. This section gives an overview of research relating to these two techniques, which will be implemented on documents from the *Jane Addams Papers Project*. The information discussed in this section, including gaps in the current research, serves as the foundation for the work outlined in this paper.

With increasing amounts of textual data in society, machine learning techniques like text classification and topic modeling are important for making sense of that data. Machine-based methods can capture meaning from texts, uncovering patterns and revealing insight about their contents. Topic modeling can aid in the analysis of texts, serving as a "valuable tool to identify and extract latent themes or topics from a collection of documents" (Detthamrong et al., 2024, p. 449). Text classification is similarly valuable, "serving as a linchpin for the categorization and delineation of diverse content types and facilitating streamlined information retrieval" (Kholwal, 2023, p. 415). Research on both text classification and topic modeling has expanded over time, leading to valuable contributions to these fields and insight on how these methods can be used to explore texts in various domains. However, this research leaves opportunities to investigate how these two techniques can be used together to analyze texts.

Text classification studies have taken place for decades, with the first paper on automatic text classification being published in 1961 (Zhang et al., 2023). Since then, research on text classification has been conducted using a variety of different datasets and models. Much of this research focuses on single-label classification problems, in which classifiers assign only one label to texts (Shaikh et al., 2023). Kholwal (2023), for instance, used text classification to

categorize BBC news articles with one of five distinct labels. This study compared different models, including Logistic Regression, Random Forest, and K-Nearest Neighbor algorithms, for the purpose of finding a classifier that can aid in information retrieval from news documents. After evaluating the models with a variety of metrics, including accuracy, precision, recall, F1-score, and support, the results showed that the Logistic Regression and Random Forest classifiers performed best (Kholwal, 2023).

In recent years, there has been an increasing focus on multilabel text classification problems. Unlike with single-label problems, examples can have more than one label in the case of multilabel classification (Shaikh et al., 2023). According to Shaikh et al. (2023), multilabel classification "is a fast-growing field of machine learning" (p. 1). There is value in studying multilabel classification since many real-world problems involve data that can be assigned to more than one category (Yuan et al., 2024). There are a variety of multilabel classification models that have been implemented in recent studies. Shaikh et al. (2023) utilized problem transformation algorithms, including Binary Relevance, Classifier Chain, and Label Powerset, which transform multilabel tasks into single-label ones. After applying these models to five different multilabel datasets, two of which were composed of textual data, Shaikh et al. (2023) found that the Classifier Chain algorithm performed well, especially on the texts. This study was valuable in demonstrating the effectiveness of problem transformation approaches for multilabel text classification tasks.

Problem transformation algorithms have been used in other studies, such as the one conducted by Arslan and Cruz (2024). The researchers aimed to implement a multilabel classification system for a company hoping to automate the categorization of business documents. Arslan and Cruz (2024) emphasized the benefits of automated classification, stating

that a machine-based method "not only enhances the efficiency of business text classification but also minimizes the inherent risks of errors often associated with manual text classification" (p. 2). Using a dataset of almost 29,000 texts with 80 distinct labels, Arslan and Cruz (2024) implemented Binary Relevance, Classifier Chain, and Label Powerset, as well as bidirectional encoder representations from transformers (BERT), which involved fine-tuning a pre-existing BERT model for their classification task. This BERT model showed the highest performance with almost 90% accuracy, but Binary Relevance had a similar F1-score, precision, and recall and only a slightly lower accuracy (Arslan & Cruz, 2024). This study, like the work of Shaikh et al. (2023), showed the usefulness of problem transformation methods and the ability of machine learning techniques to accurately classify texts.

Machine-based methods are also able to uncover trends in large collections of texts. A number of studies demonstrate the effectiveness of topic modeling in analyzing texts by identifying themes present within them. In particular, latent Dirichlet allocation (LDA) is a topic model that is "renowned for its proficiency in generating descriptive topics" (Detthamrong et al., 2024, p. 451). LDA is widely used within topic modeling research. Detthamrong et al. (2024), for instance, used LDA to explore themes in over 8,000 documents relating to digital economy research. After preprocessing the texts and deciding on the optimal number of topics using the coherence score metric and visualizations, Detthamrong et al. (2024) identified three distinct themes of digitalization, data governance, and digital transformation. According to Detthamrong et al. (2024), policymakers and organizations can use these topics to better understand the digital economy, which underscores the benefits of using LDA to identify trends in texts.

Other studies also implement LDA to analyze texts relating to specific subjects, adding value and meaning to those particular fields. To uncover trends in teacher educator research,

Özmantar et al. (2024) employed LDA on 754 scientific publications related to teacher education. Out of the 10 topics identified by the LDA model, two were related to themes that had been previously identified and explored in teacher education research. Five topics were related to themes that had been previously implied, but never formalized. This demonstrates how topic modeling can help clearly identify trends, providing evidence of their existence in texts. The remaining three topics that resulted from the model of Özmantar et al. (2024) were new themes that had not been studied before in the field of teacher educators. Therefore, LDA topic models can highlight previously undiscovered patterns in texts, generating new insight that can be valuable for the advancement of particular fields of study.

Both topic modeling and text classification are useful for analyzing texts, and there are studies that incorporate both of these techniques. Some research focuses on using topic modeling to aid in classification tasks, especially single-label ones. In the case of Luo and Li (2014), LDA helped extract features from texts to serve as the inputs for a Support Vector Machine (SVM) classifier to predict categories for news articles from the *Reuters* and *20 Newsgroups* datasets, which are widely used for text classification problems. These features were the topic distribution for each text–the estimated percentage of the text that is made up of each topic resulting from the LDA model. Luo and Li (2014) found that the SVM model performed better using just 120 features derived from LDA than it did using 1,100 features created from Document Frequency or 400 features obtained from Principal Component Analysis. Therefore, topic modeling can play a helpful role in classification pipelines for dimensionality reduction and feature extraction, improving classifier performance.

In other studies, topic modeling and text classification are not woven together, but are conducted separately. This was the case in the research of Wang et al. (2023), who implemented

LDA topic modeling and single-label text classification to better understand and categorize studies relating to Chinese hamster ovary cells, which are used in the pharmaceutical industry. Wang et al. (2023) had separate goals for each of the two techniques. Topic modeling was conducted to compare the resulting machine-generated topics with the manually assigned label that each document in the dataset had. Wang et al. (2023) labeled each document with the topic that made up the greatest percentage of its content, determining that the machine-generated topics correlated well with the manually assigned labels. After performing this topic modeling analysis, Wang et al. (2023) used Logistic Regression to classify the texts using the manually assigned labels with the goal of creating a more efficient system to categorize future texts. This study shows how topic modeling can be used to explore how well human-assigned categories align with those generated by a machine. However, the separate topic modeling and text classification analyses leave some questions as to how these two techniques might be combined.

In a similar study, Alamsyah and Girawan (2022) implemented topic modeling and text classification separately on text data consisting of consumer feedback for clothing companies. The goal of this research was to better understand the feedback to identify areas where companies could improve their products to reduce clothing waste. This study was significant since Alamsyah and Girawan (2022) used multilabel text classification to categorize clothing reviews into different subjects, such as materials and durability. Much of the research linking topic modeling and text classification focuses on single-label classification, meaning there are less studies that concern multilabel classification. Alamsyan and Girawan (2022) then used topic modeling to extract 10 distinct themes from the collection of customer reviews. Both the topic modeling and multilabel classification allowed for insight to be gained about the reviews, including the trends and common categories within them.

Other areas of research involving both multilabel text classification and topic modeling utilize these two techniques in a manner similar to that of Luo and Li (2014), namely, incorporating topic modeling into a classification pipeline. In their research on multilabel text classification, Tandon and Chatterjee (2022) proposed a new algorithm involving clustering texts and assigning labels fuzzy memberships to these clusters. The researchers experimented with different feature extraction techniques to use with this algorithm, including document embeddings, TF-IDF, fuzzy c-means clustering, and two topic models: LDA and Contextual Topic Modeling. The topic modeling and fuzzy c-means clustering methods resulted in the best classification performance (Tandon & Chatterjee, 2022). This emphasizes the conclusion of Luo and Li (2014): that topic modeling can aid in text classification tasks by extracting meaningful, useful features.

Researchers have intersected topic modeling and text classification to varying degrees. Most studies that combine these techniques investigate how topic modeling can improve classification performance. Other studies treat topic modeling and classification as separate tasks, deriving different insights that may contribute to a more general understanding of the content of texts. This opens up opportunities to explore different ways to bridge topic modeling and text classification besides using one technique to improve the other's performance. When combined in different ways, topic modeling and classification could produce new insights and introduce new pathways for the analysis of texts. Using multilabel classification in a study of this kind enhances its relevance since this field of machine learning is steadily gaining more attention.

The research on topic modeling and multilabel text classification shows how these two methods can be used to analyze and organize texts. The *Jane Addams Papers Project* Digital

Edition served as an ideal source of data on which to implement these two techniques. Topic modeling is useful for uncovering themes and insight from large collections of texts, like the one contained in the Digital Edition. Multilabel classification can be used to automate the categorization of texts, making this an appropriate method to apply to the documents from the Digital Edition since most come with several tags. By implementing both topic modeling and multilabel classification on documents from the Digital Edition, it was possible to explore how these two methods can work together to analyze texts.

Methodology

This section describes the processes of conducting topic modeling and multilabel classification, as well as the steps taken to compare the results of both techniques. Before performing any modeling, the documents and their associated tags from the *Jane Addams Papers Project* Digital Edition were examined and preprocessed. Various latent Dirichlet allocation (LDA) topic models and multilabel classifiers were then constructed for the purposes of extracting themes from the documents and predicting the tags that should be assigned to each of them. To establish a link between the topic modeling and multilabel classification, each document and tag was assigned to a topic. A series of calculations were made to quantify the overlap between tags and topics, allowing for an exploration of how topic modeling and multilabel classification relate to one another.

3.1 Data

The data consists of documents from the Digital Edition of the *Jane Addams Papers Project*. The Digital Edition contains over 20,000 texts. The transcripts of these texts are available, as well as images of the original documents. There is a variety of information associated with each text, including the title, creator, date of creation, source, a description of the text, and a list of subjects corresponding to the text's content. In addition, there are a number of labels assigned to each text to create a system of categorization. These labels, known as tags, are "used for large subject-based divisions in documents" and "allow users to look at subsections of the digital edition with ease" ("Tags," n.d.). An image of one of the documents contained in the Digital Edition, as well as its corresponding information, is shown in Figure 1.

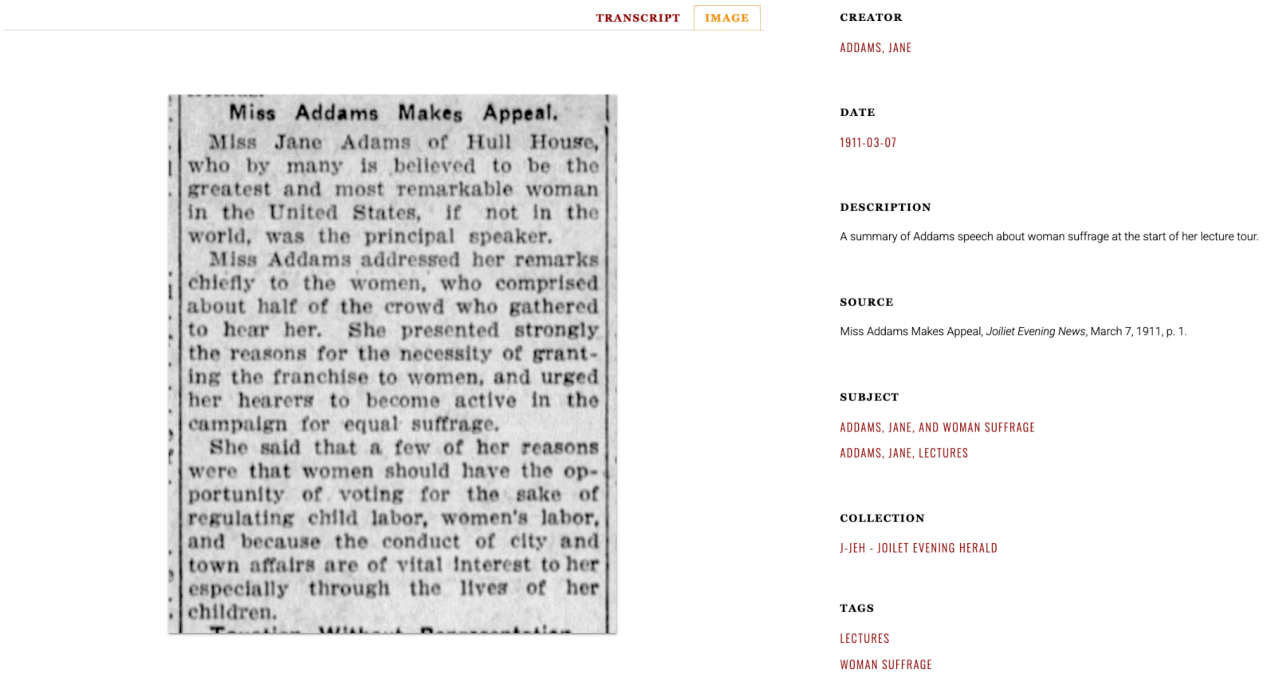


Figure 1. Example of a document and its associated metadata found in the *Jane Addams Papers Project* Digital Edition

According to Dr. Hajo, the director of the *Jane Addams Papers Project* at Ramapo College, *Project* staff members assign tags to documents by hand after reviewing their contents. There is a list of tags from which staff members can choose when labeling texts. This list was first assembled by compiling a list of common topics associated with Jane Addams. Tags can be added to the list if staff members feel none of the current ones adequately describe a document's content. Tags can also be removed from the list if they are not being used. These tags and the contents of the texts were the most relevant pieces of information for the multilabel classification and topic modeling analyses that were conducted in this study.

The subset of documents from the Digital Edition that were used in this project were speeches and articles written by Jane Addams. Since these are historical documents that have been preserved, they might not exist in the same format as when they were first written. Some of

the documents, for instance, are excerpts rather than full texts. Therefore, the results that come from using these documents are dependent on how they were preserved. In addition, the Digital Edition does not contain every speech or article written by Jane Addams, so the results also depend on which documents have been preserved. As of the time this project began, there were 835 speeches and 379 articles written by Jane Addams in the Digital Edition, creating a total of 1,214 documents. The transcripts of these documents were downloaded, along with the associated metadata for each one, including the title of the document, the date it was created, the language in which it is written, and the collection of papers from which it came. In addition, the tags that *Project* staff members assigned to each text to categorize them were also downloaded. For the purposes of this research, the data that was used were the actual text of the documents and their associated tags.

Before conducting any modeling, some data preprocessing was required. There were eight speeches and four articles missing tags, so these were removed from the dataset. Four speeches and three articles were missing their text transcriptions, so these were removed as well. Eight documents were listed as being written in a language other than English. Four of these were written in French, German, or Spanish, so they were removed from the dataset in order to ensure that all documents had a consistent language. The other four of these documents did not have a language listed, but examining the text revealed that they were written in English. Therefore, they were kept in the dataset.

Certain tags were removed from the dataset as well. Many speeches had a “Lectures” tag that reflects the document’s status as a speech rather than the content of the document. Therefore, this tag was removed since it would not provide much meaning in the analysis of the documents and their content. Similarly, many articles were assigned an “Articles” tag or a “Writings” tag,

which do not provide much information about the content of the documents themselves and were also removed. Once these tags were removed, a check was performed to determine whether there were any documents that no longer had any tags. There were five speeches that had "Lectures" as their sole tag, so these speeches were removed. After this data preprocessing, there were 815 speeches and 371 articles left in the dataset. This made a total of 1,186 documents to be used for the topic modeling and multilabel classification.

3.1.1 Tags

An examination of the tags was conducted to gain more information about them. There were a total of 201 unique tags in the dataset. Some tags occurred many times, such as "Peace," which was assigned to a total of 215 documents. Other tags were very infrequent, sometimes occurring only once in the dataset. Figure 2 shows the top 10 tags assigned to documents in the dataset. These tags are unsurprising given Jane Addams' background in social reform and her involvement in the peace, women's rights, and child labor movements during her lifetime.

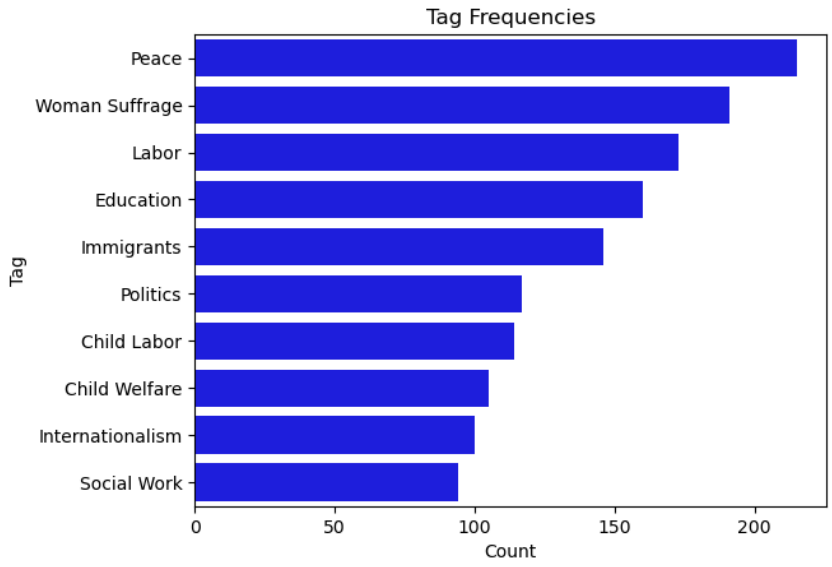


Figure 2. Top 10 tags and their frequencies

The average number of tags assigned to a document was 3.63, meaning that documents commonly have three to four tags. The minimum number of tags any document had was one, while the maximum number was 23. This high number of tags did not occur frequently, as shown in Figure 3. The majority of documents had five or fewer tags. Notice that there are no documents with zero tags, as these documents were removed.

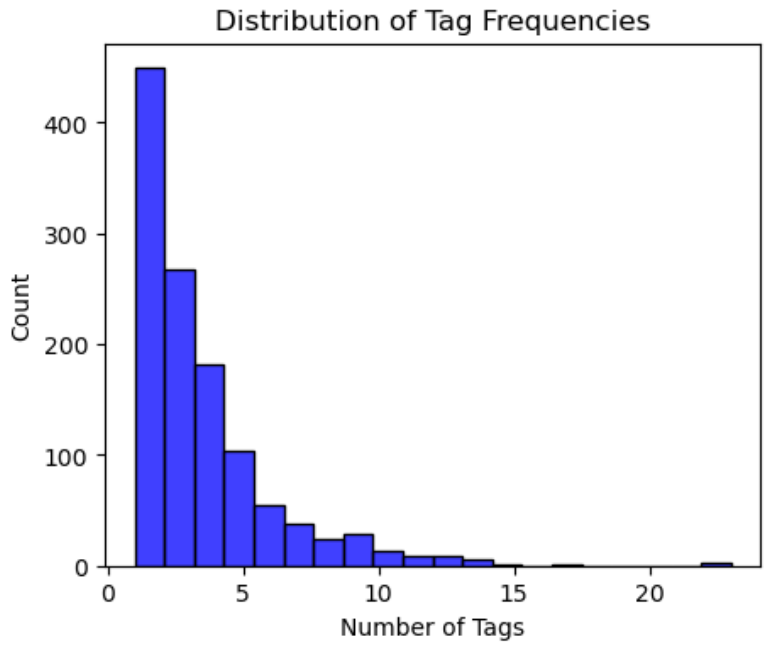


Figure 3. Histogram of tag counts

There were a total of 905 unique tag combinations in the dataset, meaning that out of the 1,186 documents, some had the same tag label combinations. However, the tag combinations were unique for a majority of documents. As seen in Figure 4, which shows pairs of tags that are often assigned to the same document, tags occurring frequently together seem related to one another. The tag "Peace," for instance, is linked to tags relating to war and international affairs, reflecting how Addams likely advocated for peace in many of her speeches and articles

concerning world conflicts. Other tag pairs, such as “Child labor, Education” and “Politics, Woman Suffrage” reflect Addams’ social reform efforts.

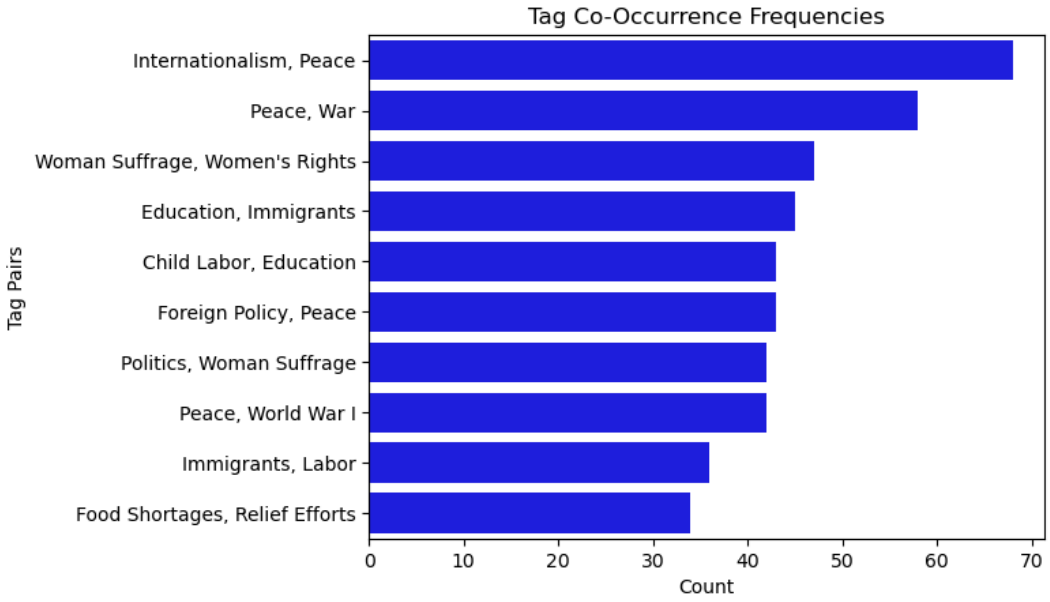


Figure 4. Top 10 tag pairs and their frequencies

3.1.2 Text

Additional preprocessing was required to prepare the content of each document for topic modeling and multilabel classification. The following steps were performed for each document: lowercasing the text, removing any characters other than letters, tokenizing the text (breaking it down into individual words), and normalizing the text through lemmatization (reducing the words to their root forms) (Kholwal, 2023; Zadgaonkar & Agrawal, 2024). Stopwords, or common words that do not offer much meaning in textual analysis like “the” and “is,” were also removed from the text using the list of English stopwords from the natural language processing package for Python, NLTK (Zadgaonkar & Agrawal, 2024). This list of stopwords was extended to include words that appeared frequently within the *Jane Addams Papers Project* documents,

but that did not contribute much meaning, including “jane,” “addams,” “mr,” and “mrs.” Some of the most frequent words are shown in Figure 5. “Woman” occurs most often across all of the documents, followed by “child,” “people,” and “men,” suggesting a focus on society and its members throughout Addams’ speeches and articles. Other frequent words are unsurprising, such as “chicago,” the site of the settlement house Addams’ founded, and “war,” which reflects Addams’ concern with world conflicts.

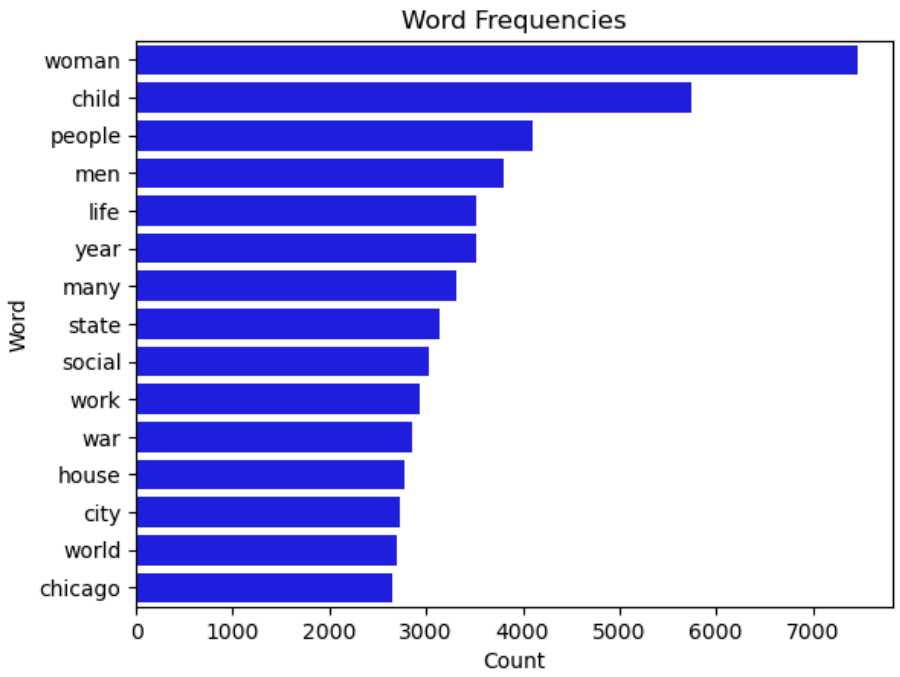


Figure 5. Top 15 words and their frequencies

3.2 Topic Modeling

Latent Dirichlet allocation (LDA) topic modeling was conducted to extract a set of topics, or themes, from the *Jane Addams Papers Project* documents. LDA is a probabilistic model that aims to predict the topics present in texts (Özmantar et al., 2024). The main assumption of LDA is that documents are created from mixtures of topics, which are themselves made up of mixtures

of words (Pavlinek & Podgorelec, 2017). More specifically, LDA assumes that a document is generated by first sampling topic probabilities from a Dirichlet distribution and randomly choosing topics from this distribution. Then, words are randomly chosen from these selected topics according to the multinomial distribution associated with each topic. The goal of LDA is to reverse this document generation process and uncover the topics making up a collection of texts (Pavlinek & Podgorelec, 2017). LDA returns two outputs: the distribution of words for each topic and the distribution of topics for each document. Topics are represented by words from the corpus of documents, each with their own associated probability, which is the likelihood of a topic generating that word. Documents are represented by each of the topics from the model, which also have their own associated probability, indicating the estimated percentage of the document that was generated by each topic (Voskergian et al., 2024).

LDA topic modeling was implemented on the version of the *Jane Addams Papers Project* documents that were tokenized and lemmatized, with stopwords and special characters removed (cleaned version). The remaining words in the cleaned documents served as the dictionary of this corpus. The texts were converted into a bag-of-words format, meaning each text was represented by the words appearing in it and the number of times those words appeared (Zadgaonkar & Agrawal, 2024). Both the dictionary and bag-of-words model were created using the Gensim library made for topic modeling in Python.

Once the documents were prepared for topic modeling, various LDA models were run with Gensim's LDA model function using different values for the three hyperparameters that have to be specified: the number of topics, alpha, and beta. Alpha corresponds to the distribution of topics per document; higher values for alpha mean that the documents will be made up of more topics (Zadgaonkar & Agrawal, 2024). On the other hand, beta corresponds to the

distribution of words per topic; higher values for beta mean that each topic will be composed of more words (Zadgaonkar & Agrawal, 2024). One method to evaluate the chosen number of topics and values of alpha and beta is to examine the model's coherence score. This metric measures the similarity of words within a topic, with higher values indicating that topics are more coherent and interpretable (Lee et al., 2024). Hyperparameter tuning was used to assess which model had the highest coherence score out of models with 10, 20, or 30 topics and alpha and beta values of 0.1, 0.5, or 0.9, which were chosen based off of the decision of Dinsa et al. (2024) to use 0.1 for both alpha and beta for their topic model. The resulting topic models produced many similar topics that were not readily interpretable.

To try and improve the results, phrase modeling was conducted on the cleaned texts. Phrase modeling detects words that frequently co-occur consecutively within a collection of texts (Zadgaonkar & Agrawal, 2024). Bigram and trigram models, for instance, identify sets of two and three words, respectively, that co-occur often and combine them into a single token, or word (Zadgaonkar & Agrawal, 2024). Gensim's Phrases model was used to create bigrams, which were added to the dictionary of words to be transformed into the bag-of-words model. In addition, words appearing in less than 10 texts or in more than 90% of the texts were filtered out of the dictionary. This decision was based on the method of Wang et al. (2023), who removed words appearing in less than five documents. Hyperparameter tuning was again conducted to test different topic numbers: 10, 20, 30, 40, or 50. This time, the alpha and beta values were not specified, instead using Gensim's default values for both, which are one divided by the number of topics ("models.lda - Latent Dirichlet Allocation," n.d.). Each of the resulting topic models again produced many overlapping topics consisting of the same words. In addition, for most of the models, the coefficients for many words throughout the topics, which indicate the probability

of a word belonging to a given topic, were zero. Therefore, these topics and the words they contained did not offer any meaning.

In a further effort to improve results, the alpha value was set to 50 divided by the number of topics, and the beta value was set to 0.01. This followed the method of Özmantar et al. (2024), who stated that these values are commonly chosen for alpha and beta. LDA topic models with these alpha and beta values were run using topic numbers of 10, 15, 20, 25, 30, 35, 40, 45, and 50 on the dictionary that included bigrams and had words occurring in less than 10 documents or more than 90% of documents removed. These topic models showed better results than the previous models. The topics were more readily interpretable and distinctive. For each model, there were no longer any topics containing words with coefficients of zero. Since these models had better results, they formed the set from which a final topic model would be chosen.

In order to evaluate these topic models with topic numbers ranging from 10 to 50, a combination of quantitative metrics, visual interpretation, and domain knowledge was used. Coherence scores were used to assess the coherence of the topics in each model. The Python library pyLDAvis for topic model visualization was also used to provide a visual representation of each model and determine whether there was overlap among the topics. With the aid of Dr. Hajo and the Assistant Editor of the *Jane Addams Papers Project*, Caitlin Biebrich, the model with 15 topics was chosen as the most suitable one based on the resulting topics and words.

3.3 Multilabel Text Classification

In addition to implementing LDA topic modeling on the *Jane Addams Papers Project* documents, multilabel classification was conducted to predict the tags associated with each document. Multilabel text classification requires preprocessing steps similar to those necessary for topic modeling, including lowercasing or uppercasing, tokenizing, and normalizing the text

(Kholwal, 2023). In addition, stopwords should be removed since they do not contribute much meaning for distinguishing between different classes of texts (Kholwal, 2023). Therefore, the cleaned version of the *Jane Addams Papers Project* texts were the features for the multilabel classifiers since they were converted to lowercase, tokenized, lemmatized, and cleared of any special characters or stopwords. The cleaned texts were then vectorized, or transformed into numerical arrays that computers can understand. A common text representation is the bag-of-words model, which transforms texts into an array that counts how many times each word in the corpus of texts occurs in each document (Özmantar et al., 2024). The bag-of-words model was implemented using CountVectorizer from Python's machine learning library, Scikit-Learn. This vectorized version of the cleaned documents was the input for the multilabel classification algorithms.

The tags assigned to the documents were the labels, or target variables, that the multilabel classifiers predicted. Some preprocessing was required to transform the set of labels assigned to each document. Using the MultiLabelBinarizer from Scikit-Learn, the labels and their associated documents were transformed into a binary matrix. Each row was a document, each column was a tag, and each cell contained either a zero or a one–a zero to indicate that the document did not have that tag and a one to indicate that the document did have that tag ("Transforming the Prediction Target (y)"). This ensured that the labels were encoded as numerical data that multilabel classification algorithms can process.

After encoding the labels, the documents and tags were split into training and test sets. Two different splitting methods were used from the iterative-stratification package developed by Bradberry (2018), which implements the iterative stratification technique for multilabel data described by Sechidis et al. (2011). This method aims to address the issue of imbalanced data

that can arise often in multilabel datasets since there could be many different classes of label combinations that contain only a small number of samples. Sechidis et al. (2011) developed an iterative stratification algorithm that works to maintain the distribution of positive examples of each label combination when dividing multilabel datasets into subsets. This lowers the probability of producing subsets with zero positive examples for a set of labels. Sechidis et al. (2011) found that this iterative stratification algorithm works well for datasets that have a large ratio of label combinations relative to examples. This is the case for the *Jane Addams Papers Project* dataset, which consisted of 1,186 documents and 905 unique label combinations.

The first splitting method was the MultilabelStratifiedKFolds class from the iterative-stratification package, which provides indices to divide the data into train and test sets. This method aims to preserve the proportion of samples of each class of labels as much as possible in each data fold (Bradberry, 2018). It is not possible to specify a desired test set size with this method, which resulted in a training set consisting of 580 documents and a test set consisting of 606 documents. Since it is not necessarily desirable to have a larger test set, a second method from the iterative-stratification package was used, the MultilabelStratifiedShuffleSplit class. This method provides indices for randomized stratified train and test sets, so that the percentage of samples of each label class is approximately the same in each split (Bradberry, 2018). With this method, it is possible to specify a desired test set size, so a 70-30% train-test split was made. This resulted in a training set size of 823 documents and a test set size of 363 documents.

After splitting the data, a number of different classification algorithms were trained on the training documents and labels. Three different problem transformation algorithms designed for multilabel data were implemented: Binary Relevance, Classifier Chain, and Label Powerset.

Problem transformation methods convert multilabel tasks into single-label ones (Shaikh et al., 2023). In the case of the Label Powerset algorithm, each unique combination of labels is treated as a separate class, and a classifier is trained to predict one of these classes, transforming multilabel classification into multiclass classification (Arslan & Cruz, 2024). With Binary Relevance, a separate binary classifier is trained for each label to predict whether a label should be assigned to a given text (Arslan & Cruz, 2024). Classifier Chain is similar to Binary Relevance, except that it forms a chain of binary classifiers, meaning that it considers the input text, as well as the output of previous classifiers, when predicting whether a label should be assigned to the text (Arslan & Cruz, 2024).

Since Binary Relevance, Classifier Chain, and Label Powerset require a base classifier, three different kinds were used in conjunction with each problem transformation method: Decision Tree, Random Forest, and Multinomial Naive Bayes. Decision Tree classifiers are tree-like models that learn decision rules from data to sort new examples into different classes ("Decision Trees," n.d.). Random Forests use ensembles of Decision Trees to make predictions, combining the output of the individual trees that are trained on subsets of the data ("RandomForestClassifier," n.d.). Multinomial Naive Bayes classifiers are commonly used for text classification and make predictions based on the assumption that features are independent from one another ("Naive Bayes," n.d.). Using these three base classifiers along with each of the three problem transformation methods resulted in nine models for the two splitting methods that were used, MultilabelStratifiedKFolds and MultilabelStratifiedShuffleSplit. Therefore, there were a total of 18 different models.

Different hyperparameter values were tested for each model with the goal of maximizing accuracy. For the Decision Tree classifiers, four, five, and six were tested for the maximum depth

value and two, four, and six were tested for the minimum number of samples required at a leaf node. For the Random Forest classifiers, the number of estimators was set to 100 or 300 and the maximum number of leaf nodes was set to 10 or 15. Finally, the values of 0.1, 0.3, 0.5, and 1.0 were tested for the alpha hyperparameter for the Multinomial Naive Bayes classifier.

To evaluate the models after they were trained, a number of metrics were used, including accuracy, precision, recall, and the F1 score. Accuracy is the fraction of correct predictions (Arslan & Cruz, 2024). The accuracy scores for both the test and training sets were calculated to look for evidence of overfitting. Models with much higher training accuracy than test accuracy might have overfit to the training data, meaning they may not generalize well to new data. Precision is the fraction of positive predictions that are correct out of all positive predictions, while recall is the fraction of positive predictions that are correct out of all actual positive examples in the data (Arslan & Cruz, 2024). The F1 score is the harmonic mean of precision and recall (Arslan & Cruz, 2024). Higher values of these metrics indicate better model performance. Both the micro-average approach, which aggregates the number of true positives, false positives, and false negatives to compute the final scores, and the macro-average approach, which calculates the average of the scores for each class, were used to calculate the precision, recall, and F1 score (Ploomber, 2023). Macro-averaging treats all classes equally, while micro-averaging gives all examples the same weight. In the case of imbalanced data, macro-averaging may make classifier performance seem worse, whereas micro-averaging can inflate a classifier's performance (Ploomber, 2023). Both averaging methods were chosen when calculating the precision, recall, and F1 score to provide a comprehensive evaluation of the models. Another metric that was used to evaluate the classifiers was the Hamming loss, which is

the fraction of labels that a classifier predicts incorrectly (Shaikh et al., 2023). In this case, lower scores are more desirable.

In addition, a metric similar to a Hamming score was utilized to choose a final classifier. A standard accuracy score can be a harsh metric for multilabel classification tasks since it considers only predicted label sets that exactly match the actual label sets for an example to be accurate ("An Introduction to Multilabel Classification," 2020). For instance, if a classifier accurately predicts two out of three labels, this classification would still be considered incorrect because the third label was not correctly predicted. In contrast to accuracy, the Hamming score considers individual correct predictions, calculating the proportion of labels that are correct (Fujishiro et al., 2023). To implement a Hamming score metric, a function was built that calculates the number of correct labels out of the total number of actual and predicted labels for each document. This translates to calculating the number of items in the intersection of the actual and predicted label sets (the number of tags that overlap) divided by the number of items in the union of the actual and predicted label sets (the total number of unique tags). This calculation was made for each document in the test set, so the average was taken to find the final Hamming score.

3.4 Examining Topics and Tags

After implementing the topic modeling and multilabel classification on the *Jane Addams Papers Project* documents, an exploration of how the tags assigned to the documents related to the topics found by LDA was conducted. As stated earlier, part of the output of the LDA topic model is the distribution of topics per document. The model returns the estimated percentage of the document that is made up of each of the topics from the model. For instance, the top two topics making up the document entitled "Woman's Contribution to the International Peace

Movement, March 30, 1934" are Topics 2 and 10, which make up about 27.84% and 23.85% of the document, respectively. The rest of the topics contribute relatively small percentages to this text. Most of the documents follow similar topic distributions as this one, with each document's most prominent topic often making up about 20% of its content. However, there are a few documents whose most prominent topics generate an estimated 50% or higher of their contents. This can be seen in Figure 6, which displays the distribution of the percentages associated with the topic that makes up the majority of each document.

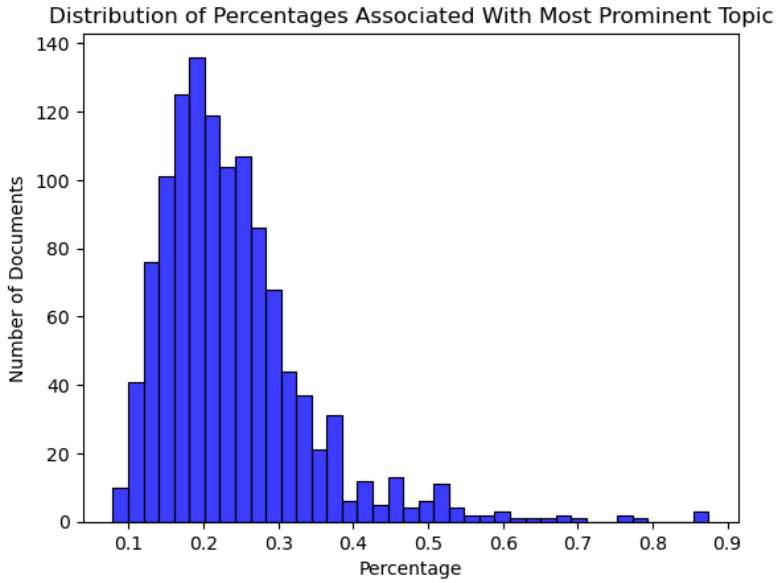


Figure 6. Frequency of documents whose leading topic generated these percentages of its content

In order to represent the documents in terms of the topics that resulted from the LDA model, each document was assigned to the topic that makes up the highest percentage of its content. This was referred to as the "dominant topic" for that document. As stated above, the most prominent topic for the document "Woman's Contribution to the International Peace Movement, March 30, 1934" was Topic 2. Therefore, Topic 2 would be this document's dominant topic.

Once the dominant topic was found for every document, an analysis comparing tags and topics was conducted. A function was designed to calculate the tag frequencies for documents that had been assigned to a particular dominant topic. These tag frequencies per topic were calculated for the actual human-labeled tags assigned to each document in the whole corpus, the actual human-labeled tags assigned to documents in the final test set, and the predicted tags assigned to documents in the final test set. The goal of this function was to determine which tags were commonly present on documents belonging to a given topic, suggesting a potential link between those tags and that topic. Comparing the actual tag frequencies and the predicted tag frequencies for documents belonging to a given topic also provided a sense of how the multilabel classifier performed in predicting tags for documents belonging to that topic.

In addition to these tag frequencies per topic, the topic distribution for each tag was determined. This was done as follows: for each tag, the total number of documents in the dataset that was labeled with that tag in the *Jane Addams Papers Project* Digital Edition was found. Then, these documents were grouped by their dominant topic. The number of documents belonging to these topic groups was divided by the total number of documents labeled with the given tag to find the percentage of documents with this tag belonging to each topic. This resulted in a breakdown of topics across tags, allowing for insight about the number and kinds of topics with which tags were associated. To demonstrate this topic distribution that was calculated for each tag, the breakdown of topics for the tag "Peace," the most common tag across documents in the dataset, is shown in Figure 7. As Figure 7 shows, out of all documents in the dataset that have been labeled with the tag "Peace" in the *Jane Addams Papers Project* Digital Edition, the greatest percentage of these documents belong to Topic 2, or have Topic 2 as their dominant topic. Documents labeled with the "Peace" tag belong to many different topics.

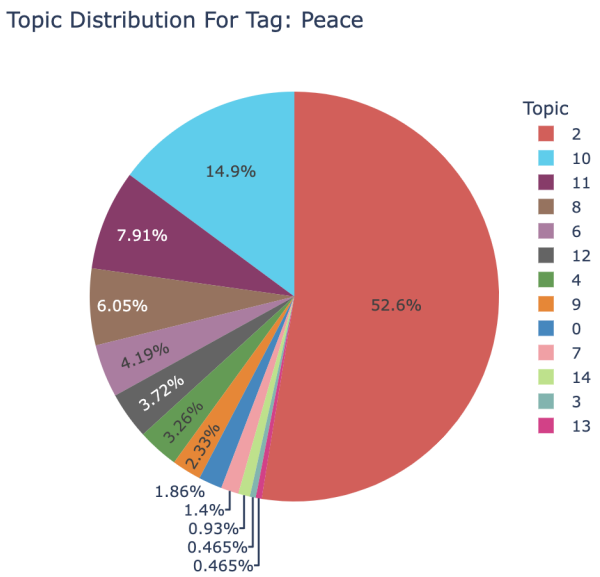


Figure 7. Pie chart showing percentage of documents labeled with "Peace" tag belonging to particular topics

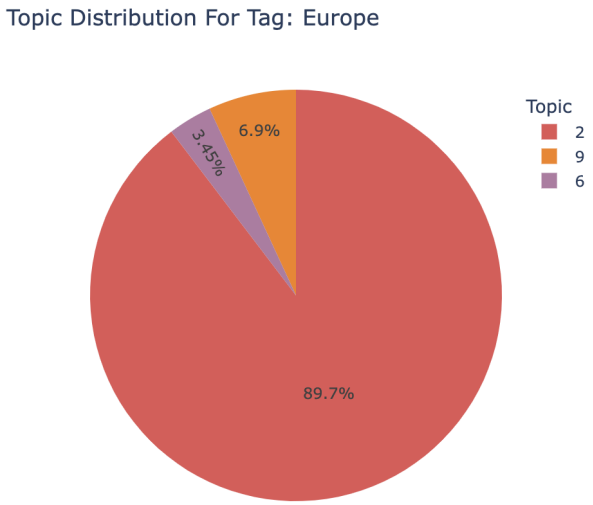


Figure 8. Pie chart showing percentage of documents labeled with "Europe" tag belonging to particular topics

In contrast, a tag like "Europe," which occurs 29 times in the dataset, can be found on documents belonging to only three topics, as shown in Figure 8. It is possible that there is less diversity in the content of documents labeled with the "Europe" tag than documents labeled with the "Peace" tag. In addition, the tag "Peace" occurs 215 times in the dataset, which is much more

often than the tag “Europe,” meaning that there is a greater chance of more topics being associated with the tag “Peace” since there are more documents with this tag to be distributed across topics.

Once this topic distribution was found for each tag, the tags were assigned to particular topics in a manner similar to how each document was assigned to its dominant topic, or the topic making up the largest estimated percentage of its content. Each tag was assigned to the topic with the highest percentage in the breakdown of topics for that tag. In other words, out of all of the documents with a given tag, the topic that was dominant for the greatest number of documents became the dominant topic for that tag. Each tag was then assigned to its dominant topic. The tag “Peace,” for instance, was assigned to Topic 2 because this topic was dominant for the greatest percentage of documents that had been labeled with “Peace” in the *Jane Addams Papers Project* Digital Edition. Looking at Figure 6, Topic 2 has the greatest segment in the pie chart showing the topic distribution for the tag “Peace,” which is why Topic 2 was considered the dominant topic for that tag. Similarly, Topic 2 was the dominant topic for the tag “Europe,” so this tag was also assigned to Topic 2. In some cases, two or more topics were dominant for an equal percentage of documents labeled with a particular tag. This was equivalent to two or more segments of the pie chart for a given tag having the same area. When this occurred, rather than arbitrarily choosing one of these topics for that tag, the topics that were dominant for equal percentages of documents were all assigned as the dominant topics for that tag. Therefore, some tags were assigned to more than one topic. Assigning each tag to its dominant topic resulted in a list of tags associated with each topic, which will be referred to as the topic tags for that topic. A full list of these topic tags can be seen in Table 7 in the Appendix.

The motivation behind assigning tags to their dominant topics and creating these lists of topic tags was to determine whether there was any degree of overlap between the tags associated with a document's dominant topic and the tags that the multilabel classifier predicted that document should have. In other words, the goal was to explore any overlap between the kinds of insights offered by performing topic modeling and classification on the documents. To quantify this overlap, a metric similar to the Hamming score described above was created. The Hamming score calculates the number of tags that occur in both the actual and predicted label sets (the intersection of the two sets) divided by the number of unique tags across the actual and predicted label sets (the union of the two sets), and computes the average of these values across all of the documents to get the final score. The adjusted Hamming score metric, which will be referred to as the topic-tag overlap, was used to compare the predicted label sets for each document with the topic tags for that document's dominant topic. This topic-tag overlap metric was applied to all of the documents in the test set, as well as to subsets of the test set consisting of only documents belonging to each one of the 15 topics from the LDA model.

The topic-tag overlap, like the Hamming score, finds the intersection of two sets: the number of tags that occur in both the predicted tag set for a document and the topic tag set associated with the document's dominant topic. However, unlike the Hamming score, the topic-tag overlap metric does not divide the number of tags that occur in both sets by the union of the two sets, but by the total number of predicted tags. The topic-tag overlap metric was designed this way because some topics are associated with many tags. The number of tags assigned to Topic 2, for instance, is 47. The multilabel classifier is unlikely to predict that many tags for a document. As a result, when trying to compute the overlap between the predicted tag sets and the topic tag sets for documents belonging to Topic 2, if the denominator of the topic-tag

overlap metric were to be the total number of unique tags in both sets (the union), the score for each document would most likely be very low. Therefore, the topic-tag overlap finds the number of tags common to both the predicted tag set and the topic tag set divided by the number of predicted tags for each document and computes the average to obtain the final score. The topic-tag overlap can be interpreted as the percentage of tags that belong to the same topic as their corresponding documents, on average.

In addition to comparing the predicted tag sets and the topic tag sets with the topic-tag overlap, the actual tag sets (the labels assigned to the documents in the *Jane Addams Papers Project* Digital Edition) and the topic tag sets were compared using this metric. This was again done for all of the documents in the test set, as well as to subsets of the test set consisting of only documents belonging to a given topic. The purpose of this comparison was to determine what percentage of actual tags assigned to a document also appeared in the topic tags associated with the document's dominant topic, on average. This would help uncover whether there was more similarity between the topic tags associated with documents and the labels that were predicted for them versus the labels that were actually assigned to them.

Finally, the original Hamming score metric was used to compare the actual tag sets and the predicted tag sets for the whole test set, as well as subsets of the test set consisting of only documents belonging to a given topic. This helped determine whether the multilabel classifier performed better or worse in predicting tags for test documents belonging to particular topics than for all of the documents in the test set.

To summarize, three different computations were made on the test set of documents and 15 subsets of the test set made up of documents belonging to one of the 15 topics produced from the LDA topic model. These computations were: (1) using the Hamming score, the percentage of

tags that were correctly predicted for each document, on average, (2) using the tag-topic overlap, the percentage of each document's predicted tags belonging to that document's dominant topic, on average, and (3) using the tag-topic overlap, the percentage of each document's actual tags belonging to that document's dominant topic, on average.

The Hamming score and tag-topic overlap calculations were intended to bridge the results of topic modeling and multilabel classification after the most suitable models were chosen. Various steps were taken to improve model performance. For the LDA topic model, a range of topic numbers was tested, as well as different values for alpha and beta. Hyperparameter tuning was also conducted for the multilabel classifiers, and two different splitting methods were used to divide the documents into train and test sets. Implementing topic modeling and multilabel classification allowed for an analysis of the topics and tags associated with Jane Addams' speeches and articles, which will be discussed in the next section.

Analysis and Discussion

Once the various LDA topic models and multilabel classifiers were constructed, they were evaluated to determine the optimal topic model and classifier. This section discusses the insights that the topic model captured from Jane Addams' speeches and articles, as well as the multilabel classifier's performance in predicting tags for the documents. An analysis of the topics that make up each document and their assigned tags revealed that there is a connection between the information that topic modeling and multilabel classification convey about texts. This connection was explored more deeply through an examination of three specific topics.

4.1 Topic Modeling Results

The topic model with 15 topics was chosen as the best model after comparing this one to those with 10, 20, 25, 30, 35, 40, 45, and 50 topics. Coherence scores, visual interpretation, and domain knowledge were all considered when analyzing the different models. The coherence scores of the models, shown in Figure 9, were relatively similar, ranging from about 0.34 to 0.40. Higher coherence scores are more desirable, indicating that the topics have better quality and that the words comprising each topic relate to one another better (Lee et al., 2024). The models with 10, 15, and 20 topics had the highest coherence scores. As shown in Figure 9, the coherence scores began to steadily decline past 30 topics.

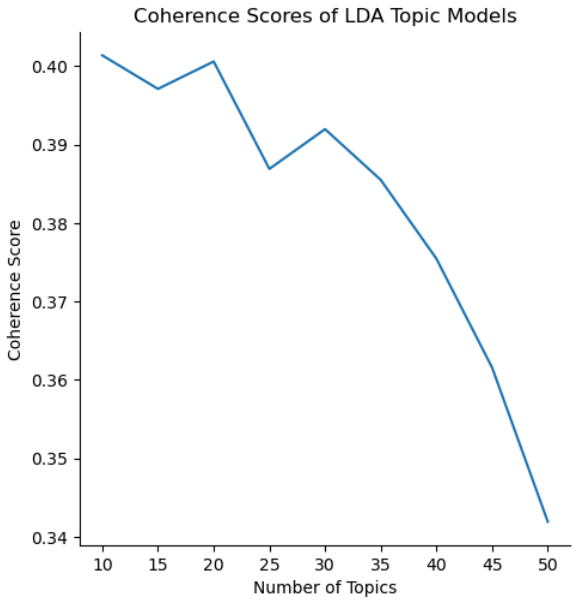


Figure 9. Coherence scores for topic models with different topic numbers

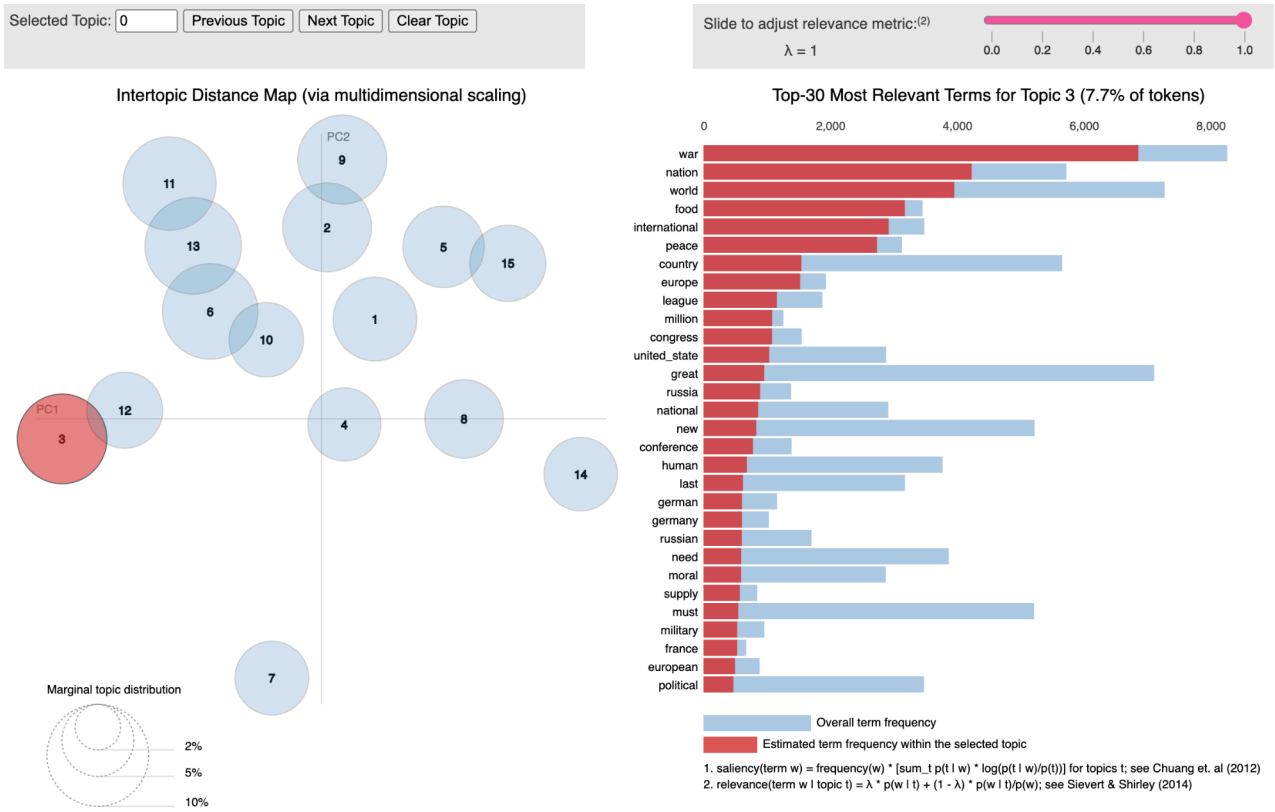


Figure 10. pyLDAvis output for model with 15 topics

Examining visual representations of the topic models showed further evidence of the models with 10, 15, or 20 topics being more suitable. The visualizations provided by the pyLDAvis library are useful tools for analyzing topic models. Figure 10 shows the visualization produced by pyLDAvis for the model with 15 topics. The left side of the visualization shows the topics represented by circles, arranged in two-dimensional space. Topics located closer to one another are more similar, and the larger the circle, the more prevalent that topic is within the corpus of documents. Hovering over a particular circle reveals the words associated with that topic on the right side of the visualization. In Figure 10, Topic 3 is selected. This is actually Topic 2 from the LDA topic model with 15 topics. The pyLDAvis library uses the numbers 1 through 15 to label topics, whereas the LDA model uses the numbers 0 through 14. Therefore, the pyLDAvis visualization increases the numbers labeling each topic from the LDA model by one. The blue bars next to each word indicate the frequency of that word across all of the documents. The red bars show the estimated number of times the selected topic generated each word within the documents. If the red bar is almost the same length of the blue bar for a given word, it means that the word appears almost exclusively in documents belonging to the selected topic (Wang et al., 2023).

The pyLDAvis visualizations revealed that the models with high numbers of topics resulted in topics that overlapped more. Figure 11 shows the pyLDAvis output for the model with 20 topics. Compared to the visualization in Figure 10, which shows the model with 15 topics, the left side of the output shown in Figure 11 displays a greater overlap among the topics, especially in the top right corner. Greater overlap means there could be some redundancy between the topics, suggesting that there might be too many of them. The pyLDAvis visualizations showed that as the number of topics increased, so too did the amount of overlap.

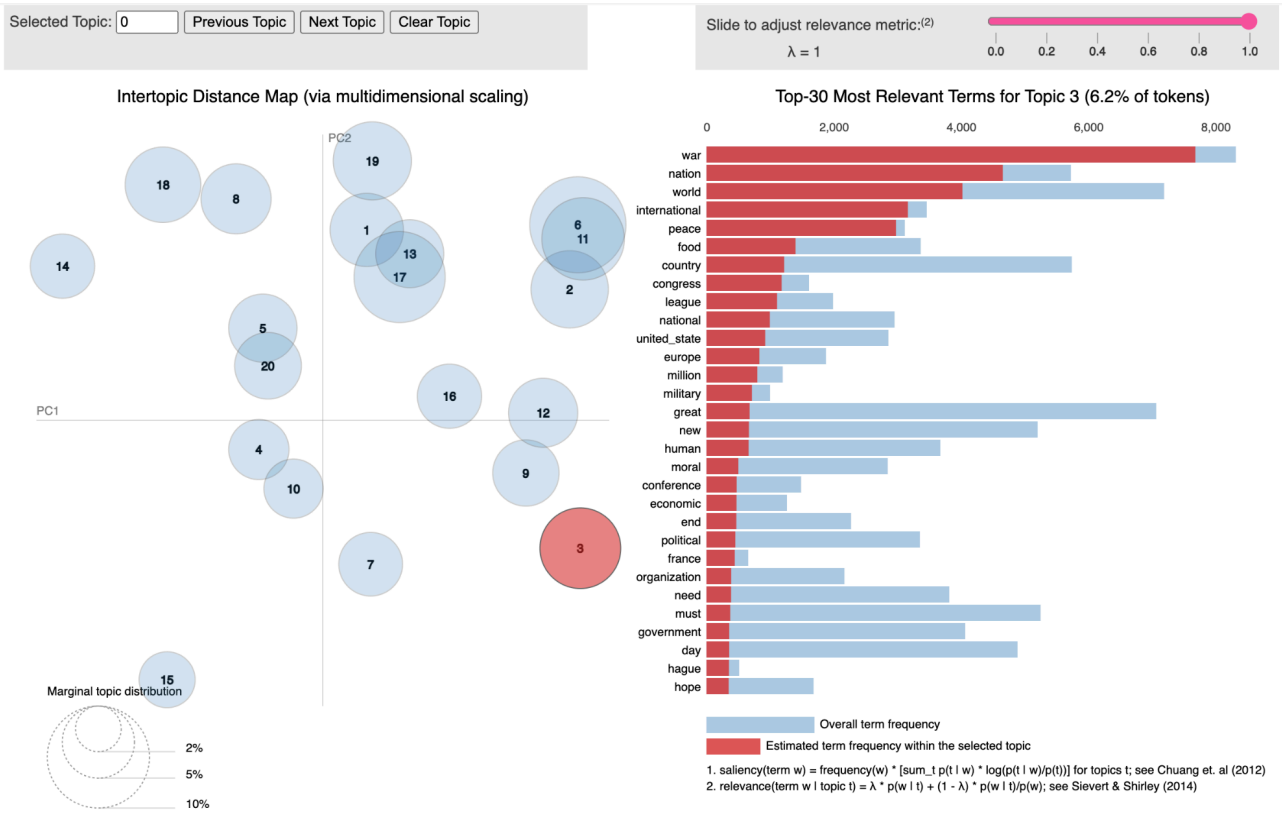


Figure 11. pyLDAvis output for model with 20 topics

Since higher numbers of topics resulted in greater overlap, the set of candidates for the final topic model was narrowed down to the models with 10, 15, and 20 topics. Using their knowledge of Jane Addams, Dr. Hajo and the Assistant Editor, Ms. Biebrich, reviewed the resulting topics and aided in choosing a final model. There was some similarity across the topics in the model with 20 of them, suggesting the topics lacked some distinctiveness. A few of the topics appeared to lack meaning as well; it was difficult to discern a theme from the words associated with these topics. The model with 10 topics, on the other hand, seemed to lack some of the nuance and meaning that models with more topics were able to capture. Therefore, the model with 15 topics was chosen as the one most suitable for the documents from the *Jane Addams Papers Project*. This model produced mostly distinctive topics that were able to capture more nuanced themes within the texts. The 15 topics from this model are shown in Table 1. Each

topic is represented by a set of words and associated probabilities. Since LDA assumes texts are generated by randomly choosing a topic and then randomly choosing a word within that topic, the word probabilities indicate how likely it is that a given word will be chosen once a topic is selected. For instance, Topic 2 has the highest probability of generating the word “war.”

Topic Number	Topic Words
0	0.031*"girl" + 0.017*"men" + 0.014*"law" + 0.012*"state" + 0.011*"young" + 0.011*"family" + 0.011*"life" + 0.010*"evil" + 0.009*"mother" + 0.009*"man" + 0.008*"father" + 0.008*"child" + 0.008*"social" + 0.007*"business" + 0.007*"wage"
1	0.024*"city" + 0.021*"life" + 0.014*"school" + 0.014*"industrial" + 0.012*"education" + 0.009*"make" + 0.009*"people" + 0.008*"modern" + 0.008*"play" + 0.007*"young" + 0.007*"public" + 0.007*"factory" + 0.007*"must" + 0.006*"industry" + 0.006*"yet"
2	0.054*"war" + 0.033*"nation" + 0.031*"world" + 0.025*"food" + 0.023*"international" + 0.021*"peace" + 0.012*"country" + 0.012*"europe" + 0.009*"league" + 0.008*"million" + 0.008*"congress" + 0.008*"united_state" + 0.008*"great" + 0.007*"russia" + 0.007*"national"
3	0.107*"woman" + 0.025*"social" + 0.021*"work" + 0.018*"condition" + 0.018*"college" + 0.017*"must" + 0.013*"life" + 0.010*"home" + 0.010*"problem" + 0.009*"new" + 0.009*"household" + 0.008*"service" + 0.008*"year" + 0.007*"done" + 0.007*"study"
4	0.023*"hull_house" + 0.015*"settlement" + 0.015*"house" + 0.014*"little" + 0.012*"people" + 0.011*"first" + 0.011*"day" + 0.010*"time" + 0.010*"came" + 0.009*"much" + 0.008*"neighborhood" + 0.008*"come" + 0.007*"way" + 0.007*"friend" + 0.006*"poor"
5	0.013*"many" + 0.012*"labor" + 0.012*"state" + 0.010*"men" + 0.008*"year" + 0.008*"long" + 0.007*"old" + 0.007*"first" + 0.007*"made" + 0.007*"might" + 0.007*"certainly" + 0.006*"legislation" + 0.006*"protection" + 0.006*"standard" + 0.006*"much"
6	0.132*"woman" + 0.021*"political" + 0.020*"vote" + 0.019*"men" + 0.016*"party" + 0.015*"suffrage" + 0.010*"state" + 0.010*"city" + 0.009*"municipal" + 0.009*"convention" + 0.009*"national" + 0.009*"question" + 0.009*"meeting" + 0.008*"government" + 0.008*"franchise"
7	0.026*"year" + 0.021*"public" + 0.020*"chicago" + 0.015*"committee" + 0.014*"first" + 0.014*"school" + 0.013*"child" + 0.013*"state" + 0.012*"club" + 0.011*"illinois" + 0.011*"board" + 0.011*"made" + 0.010*"member" + 0.008*"president" + 0.008*"many"

8	0.038*"people" + 0.020*"say" + 0.019*"know" + 0.017*"many" + 0.015*"think" + 0.014*"much" + 0.014*"get" + 0.013*"go" + 0.013*"something" + 0.013*"great" + 0.012*"sort" + 0.011*"course" + 0.010*"going" + 0.010*"come" + 0.010*"done"
9	0.030*"immigrant" + 0.027*"american" + 0.024*"people" + 0.018*"america" + 0.017*"italian" + 0.014*"social" + 0.013*"russian" + 0.012*"method" + 0.012*"hull_house" + 0.011*"settlement" + 0.011*"country" + 0.010*"class" + 0.010*"colony" + 0.008*"great" + 0.008*"among"
10	0.030*"men" + 0.010*"human" + 0.009*"man" + 0.009*"moment" + 0.009*"might" + 0.008*"life" + 0.008*"come" + 0.007*"country" + 0.007*"young" + 0.007*"must" + 0.007*"people" + 0.007*"say" + 0.006*"world" + 0.006*"quite" + 0.006*"another"
11	0.021*"government" + 0.017*"united_state" + 0.015*"nation" + 0.014*"world" + 0.012*"union" + 0.011*"war" + 0.010*"country" + 0.009*"court" + 0.009*"opinion" + 0.008*"organization" + 0.007*"league" + 0.007*"act" + 0.006*"international" + 0.006*"use" + 0.006*"representative"
12	0.019*"life" + 0.013*"social" + 0.011*"sense" + 0.009*"mind" + 0.008*"moral" + 0.007*"experience" + 0.007*"even" + 0.007*"man" + 0.007*"great" + 0.006*"many" + 0.006*"never" + 0.006*"new" + 0.006*"year" + 0.006*"force" + 0.006*"time"
13	0.038*"city" + 0.038*"chicago" + 0.037*"boy" + 0.017*"girl" + 0.014*"social" + 0.013*"condition" + 0.013*"police" + 0.013*"house" + 0.012*"public" + 0.011*"young" + 0.011*"home" + 0.011*"work" + 0.010*"street" + 0.010*"school" + 0.009*"juvenile_court"
14	0.143*"child" + 0.050*"work" + 0.038*"labor" + 0.019*"little" + 0.019*"factory" +0.016*"year" + 0.013*"go" + 0.011*"school" + 0.010*"see" + 0.010*"take" + 0.010*"come" + 0.009*"mother" + 0.009*"day" + 0.008*"age" + 0.008*"working"

Table 1. LDA topic modeling results for final model with 15 topics

The topic model extracted a number of themes from the *Jane Addams Papers Project* documents. With the help of Dr. Hajo and Ms. Briebich, these themes were identified to include women's suffrage (Topic 6), child labor (Topics 13 and 14), Hull-House (Topics 4 and 9), and international affairs and World War I (Topic 2). These themes are unsurprising given Jane Addams' background in social work and her participation in reform movements, such as the women's rights, settlement, and peace movements. While the themes of certain topics connect, there are slight nuances of meaning contained in the topics that differentiate them. Topics 13 and 14, for instance, both relate to child labor. Unlike Topic 14, however, Topic 13 also appears to

relate to child crime, as demonstrated by some of the words belonging to Topic 13, such as "police" and "juvenile court." This suggests that some documents might discuss similar subjects, such as children, but in different contexts, like child labor or child crime. Other topics are more distinctive, such as Topic 6, which appears to be the only topic mainly related to women's suffrage. The topics, therefore, extract some of the main subjects present in the texts from the *Jane Addams Papers Project*, while also identifying some more nuanced themes.

In addition to the word distributions for each topic, the model outputs topic distributions for each document. In other words, the model assigns a list of topics to each text and associated probabilities that estimate how prevalent that topic is within the text. For the purposes of this analysis, the topic that was most prevalent in a document was labeled as that document's dominant topic. The document was then assigned to that dominant topic. The number of documents assigned to each topic is shown in Figure 12. Topic 2, which relates to World War I,

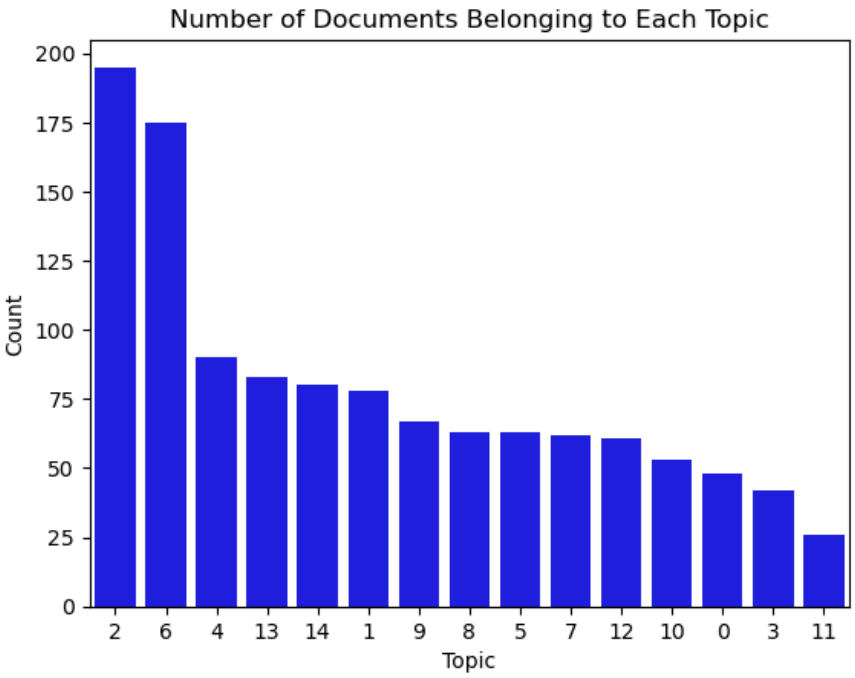


Figure 12. Frequency of documents per topic

peace, and international affairs, had the highest frequency of documents, with close to 200 texts belonging to this topic. The next most frequent topic was Topic 6, which relates to politics and women's rights, including women's suffrage. The remaining topics had relatively similar numbers of documents.

4.2 Multilabel Text Classification Results

The results of the different multilabel classification models are shown in Tables 2 and 3. In these tables, BR denotes Binary Relevance, CC denotes Classifier Chain, and LP denotes Label Powerset. Table 2 contains the evaluation metrics for the classifiers that were trained on data split using the MultilabelStratifiedKFolds method, which resulted in a train-test split of 580 training documents and 606 test documents. Table 3 contains the metrics for the classifiers trained on data split using the MultilabelStratifiedShuffleSplit method, which resulted in a train-test split of 823 training documents and 363 test documents. For each classifier, the best hyperparameters that were found during hyperparameter tuning are listed. The blue boxes in both Tables 2 and 3 mark the best value for each metric. This is the highest value in each category, except for Hamming loss, in which case lower values indicate better model performance. As can be seen from Tables 2 and 3, no single classifier performed the best in all of the evaluation categories. For both splitting methods, the Label Powerset Multinomial Naive Bayes and the Binary Relevance Random Forest had the best metrics in more than one category.

Model	Test Accuracy	Training Accuracy	Hamming Loss	Micro-Precision	Macro-Precision	Micro-Recall	Macro-Recall	Micro-F1	Macro-F1
BR Decision Tree (max depth: 4, min samples leaf: 2)	0.0545	0.2121	0.0178	0.4922	0.2447	0.3108	0.1696	0.3810	0.1895
BR Random Forest (estimators: 100, max leaf nodes: 15)	0.1320	0.2569	0.0158	0.8218	0.4155	0.1335	0.1250	0.2296	0.1738
BR Multinomial Naive Bayes (alpha: 0.3)	0.0644	0.5259	0.0206	0.4211	0.2683	0.4573	0.2293	0.4385	0.2295
CC Decision Tree (max depth: 5, min samples leaf: 2)	0.0462	0.2655	0.0184	0.4669	0.2340	0.3098	0.1681	0.3725	0.1833
CC Random Forest (estimators: 100, max leaf nodes: 10)	0.0066	0.1345	0.0161	0.8532	0.3981	0.1003	0.1038	0.1795	0.1477
CC Multinomial Naive Bayes (alpha: 0.3)	0.0611	0.5224	0.0208	0.4164	0.2549	0.4578	0.2299	0.4361	0.2227
LP Decision Tree (max depth: 5, min samples leaf: 6)	0.0314	0.0638	0.0210	0.2425	0.0366	0.0905	0.0173	0.1318	0.0168
LP Random Forest (estimators: 300, max leaf nodes: 15)	0.0776	0.5655	0.0187	0.4408	0.3932	0.2277	0.1613	0.3003	0.2912
LP Multinomial Naive Bayes (alpha: 0.3)	0.0858	0.9431	0.0335	0.2495	0.2734	0.4489	0.2485	0.3207	0.2145

Table 2. Classification results using the MultilabelStratifiedKFolds method

Model	Test Accuracy	Training Accuracy	Hamming Loss	Micro-Precision	Macro-Precision	Micro-Recall	Macro-Recall	Micro-F1	Macro-F1
BR Decision Tree (max depth: 4, min samples leaf: 4)	0.0386	0.1543	0.0175	0.5163	0.2062	0.3270	0.1495	0.4004	0.1609
BR Random Forest (estimators: 100, max leaf nodes: 15)	0.0165	0.1713	0.0160	0.8542	0.4164	0.1302	0.1265	0.2259	0.1750
BR Multinomial Naive Bayes (alpha: 0.7)	0.0634	0.2819	0.0181	0.4944	0.2435	0.4364	0.1823	0.4636	0.1958
CC Decision Tree (max depth: 4, min samples leaf: 4)	0.0523	0.1556	0.0172	0.5344	0.2136	0.3208	0.1507	0.4010	0.1656
CC Random Forest (estimators: 100, max leaf nodes: 15)	0.0193	0.1798	0.0160	0.8465	0.4120	0.1309	0.1379	0.2268	0.1853
CC Multinomial Naive Bayes (alpha: 0.7)	0.0634	0.2807	0.0181	0.4931	0.2461	0.4349	0.1820	0.4622	0.1960
LP Decision Tree (max depth: 4, min samples leaf: 2)	0.0248	0.0680	0.0206	0.2380	0.0073	0.0681	0.0093	0.1060	0.0066
LP Random Forest (estimators: 100, max leaf nodes: 15)	0.0413	0.2005	0.0205	0.2974	0.2214	0.1057	0.0601	0.1559	0.0804
LP Multinomial Naive Bayes (alpha: 0.1)	0.0854	0.9684	0.0226	0.3865	0.2970	0.4510	0.3118	0.4163	0.2778

Table 3. Classification results for the MultilabelStratifiedShuffleSplit method

To further analyze performance, the output of each classifier on the test documents was examined. This output was the tags that the classifier assigned to each document. For both splitting methods, the Binary Relevance Decision Trees and Random Forests, as well as the Classifier Chain Decision Trees and Random Forests, outputted many blank predictions, meaning that the model did not predict any tags for a document. The Label Powerset Decision Trees and Random Forests did not perform very well either, outputting the same tags for almost every document.

Therefore, the Decision Tree and Random Forest models were removed from consideration for a final classification model. This left the Multinomial Naive Bayes models. For both splitting methods, the Label Powerset Multinomial Naive Bayes classifiers showed evidence of overfitting, with much higher training accuracy scores close to 100%. The Binary Relevance Multinomial Naive Bayes and the Classifier Chain Multinomial Naive Bayes models for both splitting methods were chosen as the final four models for consideration. These four classifiers outputted blanks for some of the test documents, but for a much smaller percentage than the Decision Trees and Random Forests. For the MultilabelStratifiedKFolds method, the Binary Relevance Multinomial Naive Bayes model outputted 55 blanks, while the Classifier Chain Multinomial Naive Bayes outputted 56 blanks. Both models outputted 44 blank predictions for the MultilabelStratifiedShuffleSplit method. All four models resulted in similar accuracy scores. However, the Binary Relevance and Classifier Chain Multinomial Naive Bayes models that were trained using the MultilabelStratifiedKFolds method experienced slightly more overfitting than the models trained using the MultilabelStratifiedShuffleSplit method since their training accuracy scores were much higher compared to their test accuracies.

In addition, the Binary Relevance and Classifier Chain Multinomial Naive Bayes models that were trained using the MultilabelStratifiedKFolds method showed slightly worse performance when the Hamming score metric described earlier was used to assess the final four models. This metric, which calculates the percentage of correct labels, was designed to assess accuracy in a less harsh manner than the standard accuracy measurement. As can be seen in Tables 2 and 3, the accuracy scores for each classifier are extremely low. As stated earlier, this is because this accuracy metric only considers predicted label sets that exactly match the correct label sets to be accurate. Therefore, the Hamming score metric was used to provide a better sense of how the final four models performed since the metric accounts for individual correct predictions. The results are shown in Table 4. The Binary Relevance Multinomial Naive Bayes model trained using the MultilabelStratifiedShuffleSplit method resulted in the highest Hamming score of 0.3182. This means that on average, this model predicted about 31.82% of labels correctly. Since this model had the highest Hamming score, it was selected as the final classifier.

Model	Hamming Score
BR Multinomial Naive Bayes - KFolds	0.3084
CC Multinomial Naive Bayes - KFolds	0.3057
BR Multinomial Naive Bayes - StratifiedShuffleSplit	0.3182
CC Multinomial Naive Bayes - StratifiedShuffleSplit	0.3156

Table 4. Hamming scores for Binary Relevance (BR) and Classifier Chain (CC) Multinomial Naive Bayes models for MultilabelStratifiedKFolds and MultilabelStratifiedShuffleSplit methods

After choosing this model, more hyperparameter tuning was conducted. The alpha value was originally set to be 0.7, but decreasing alpha to be 0.2 reduced the number of blank predictions from 44 to 22 and resulted in a slightly higher Hamming score, 0.3287. To summarize, the final classifier was the Binary Relevance Multinomial Naive Bayes model trained using the MultilabelStratifiedShuffleSplit method with an alpha value of 0.2. The final metrics for this model are shown in Table 5. The test accuracy decreased slightly from 0.0634 to 0.0496, but the decision was made to change alpha from 0.7 to 0.2 since lowering alpha resulted in a higher Hamming score and less blank predictions. In addition, the values for macro-precision, micro-recall, macro-recall, micro-F1, and macro-F1 all increased when alpha was changed to 0.2, contributing to the decision to use this alpha value for the final Binary Relevance Multinomial Naive Bayes model.

Evaluation Metric	Score
Test Accuracy	0.0496
Training Accuracy	0.4058
Hamming Loss	0.0214
Micro-Precision	0.4255
Macro-Precision	0.2932
Micro-Recall	0.5513
Macro-Recall	0.3398
Micro-F1	0.4803
Macro-F1	0.2960
Hamming Score	0.3287

Table 5. Results for the final classifier: Binary Relevance Multinomial Naive Bayes trained using MultilabelStratifiedShuffleSplit with alpha set to 0.2

4.3 Analysis of Topics and Tags

To examine the performance of the Binary Relevance Multinomial Naive Bayes model in relation to the topics produced by the LDA topic model and explore any relation between the assigned tags and associated topics of documents, several metrics were calculated on the whole test set and subsets consisting of documents belonging to each one of the 15 topics. First, the Hamming score metric was used to find the percentage of tags that were correctly predicted for each document, on average. Second, the topic-tag overlap metric was used to calculate the percentage of each document's predicted tags belonging to that document's dominant topic, on average. Third, the topic-tag overlap metric was used to calculate the percentage of each document's actual tags belonging to that document's dominant topic, on average. The results for all three of these calculations are shown in Table 6.

For each document in the test set, about 32.87% of tags were correctly predicted, on average. As Table 6 shows, the percentage of tags that were correctly predicted, on average, for each document varied when the test set was restricted to only documents with a given dominant topic. For each of the documents belonging to Topic 6, for instance, about 43.24% of tags were correctly predicted, on average, which is a greater percentage than for all of the documents in the test set. Other subsets of the test set experienced lower percentages of tags that were correctly predicted, such as the documents belonging to Topics 1 and 3. This suggests that the multilabel classifier was better at predicting the tags for documents belonging to some topics over others.

Subset of Test Documents	Percent of Correctly Predicted Tags	Percent of Predicted Tags Belonging to Dominant Topic of Corresponding Documents	Percent of Actual Tags Belonging to Dominant Topic of Corresponding Documents
All test documents	32.8683	42.6972	45.1032
Topic 0	33.6954	18.8170	31.4583
Topic 1	30.7087	31.0273	33.3951
Topic 2	40.5775	92.0295	81.7269
Topic 3	26.2745	1.1765	5.5882
Topic 4	23.7029	48.6594	64.1951
Topic 5	29.6733	41.1586	45.8140
Topic 6	43.2432	80.0163	68.2240
Topic 7	30.6293	16.6179	17.7990
Topic 8	27.8470	2.2559	5.8182
Topic 9	30.6912	34.3571	43.6265
Topic 10	36.3943	2.1825	8.8955
Topic 11	31.4863	5.0215	13.2828
Topic 12	32.3528	17.3529	39.4188
Topic 13	14.8312	9.7562	29.8551
Topic 14	27.7428	45.5327	52.5183

Table 6. The three calculations made for documents in the test set and subsets of documents belonging to each one of the 15 topics: (1) for each document, the percent of tags that were correctly predicted, on average, (2) for each document, the percent of predicted tags that belong to the document's dominant topic, on average, and (3) for each document, the percent of actual tags that belong to the document's dominant topic, on average.

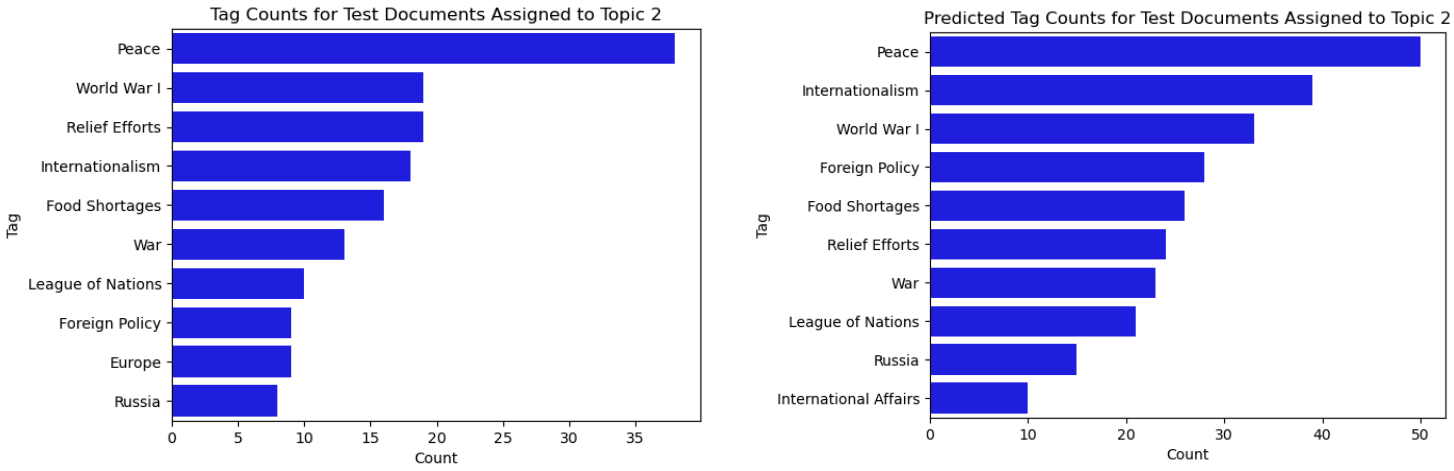
For each document in the test set, about 42.70% of the predicted tags and 45.10% of the actual tags also belonged to the list of topic tags associated with the document's dominant topic, on average. Therefore, around slightly less than half of the time, a document's actual and

predicted tags aligned with the tags belonging to that document's dominant topic. This suggests that there is some overlap between the topic and tags associated with a given document. This overlap was more apparent in documents belonging to some topics over others, however, as shown in Table 6. To explore the overlap between topics and tags and why it may have differed for documents belonging to certain topics, the following analysis will focus on three topics: 2, 11, and 14.

Topic 2 relates to peace, war, and international conflicts and affairs. The top words associated with Topic 2 are shown in Figure 13. Topic 2 was the most frequently occurring topic in the dataset; the greatest number of documents had Topic 2 as their dominant topic. As shown in Table 6, about 40.58% of tags were correctly predicted for each document belonging to Topic 2, on average. Therefore, the multilabel classifier performed better on documents belonging to Topic 2 than on all of the documents in the test set overall. Many of the most frequently

Figure 13. Words representing Topic 2, sized by associated word probabilities

occurring tags for documents belonging to Topic 2 in the test set were among the tags that were most frequently predicted for these documents, as shown in Figure 14.



(a) **(b)**

Figure 14. (a) Top 10 actual tag frequencies for test documents belonging to Topic 2 and (b) Top 10 predicted tag frequencies for test documents belonging to Topic 2

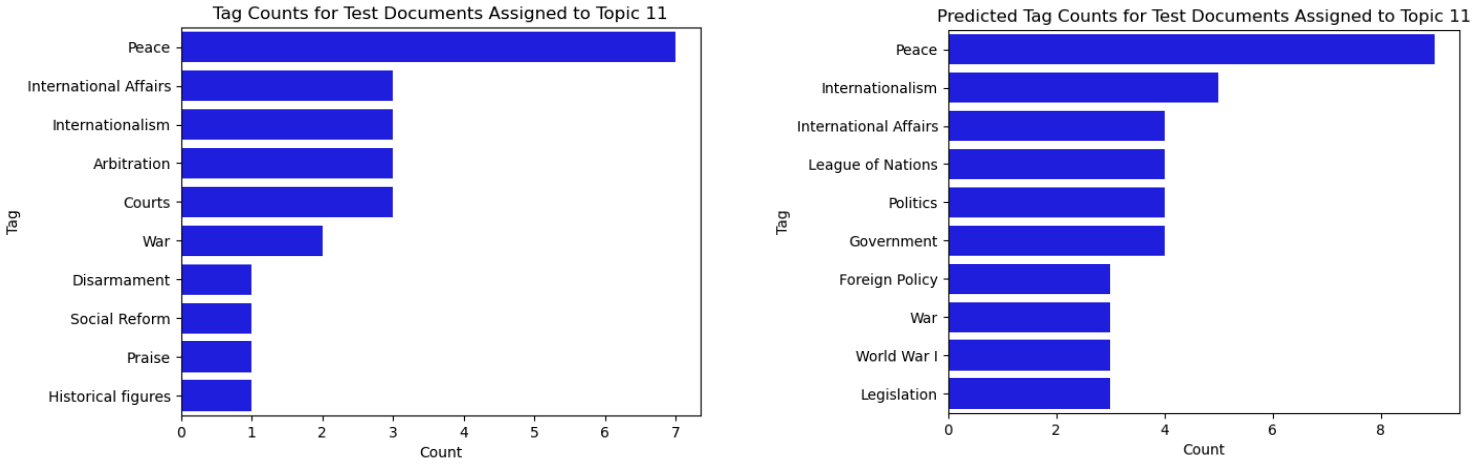
"Peace" is the most frequently occuring tag, and the multilabel classifier also predicted this tag most often. The classifier also predicted most of the other top occurring tags, such as "Internationalism" and "World War I," just with slightly different frequencies than what actually occurs among these test documents belonging to Topic 2. The most frequently occurring tags that have actually been used to label these documents relate to Topic 2's theme of war and internationalism. The same is true for the predicted tags for these documents. Therefore, both the multilabel classification model and topic model picked up on the main ideas in the texts, providing similar kinds of information. Both the topic to which these documents had been assigned–the topic making up the greatest percentage of the content of the documents–and the tags predicted to belong to these documents indicate that the content of these texts have to do with war, peace, and internationalism.

The overlap in the information provided by the multilabel classification model and the topic model for documents belonging to Topic 2 is further demonstrated by the tag-topic overlap. On average, about 92.03% of the predicted tags for the documents belonging to Topic 2 were

also in the list of tags associated with Topic 2. A full list of Topic 2's tags are shown in the Appendix, but among these tags are "World War I," "Internationalism," "International Affairs," "Peace," "Relief Efforts," "War," and "Food Shortages." These were among the top tags being predicted for documents belonging to Topic 2, as shown in Figure 14 (b). Since the documents belonging to Topic 2 were frequently assigned tags that Topic 2 itself had been assigned, this indicates further evidence of a link between a document's predicted tags and dominant topic. In the case of Topic 2, the tags that this topic suggested should belong to documents mostly related to Topic 2 are often the tags that the multilabel classifier predicted.

This is not the case for every topic, however. For the documents belonging to Topic 11, for instance, only about 5.02% of each document's predicted tags also appeared in the list of topic tags that have been assigned to Topic 11. Taking a closer look at Topic 11's word distribution, shown in Figure 15, reveals that this topic is similar to Topic 2; both relate to war and international affairs. However, Topic 11 places greater emphasis on words like "United States" and "government." Therefore, Dr. Hajo and Ms. Biebrich suggested that this topic relates to the role of the United States in international affairs, as well as government and diplomacy.

Figure 15. Words representing Topic 11, sized by associated word probabilities



(a) **(b)**

Figure 16. (a) Top 10 actual tag frequencies for test documents belonging to Topic 11 and (b) Top 10 predicted tag frequencies for test documents belonging to Topic 11

The tags assigned to documents belonging to Topics 2 and 11–both the actual and predicted tags–further demonstrate the similarities between the two topics. Figure 16 shows the most frequent tags assigned to documents in the test set belonging to Topic 11. Many of the most frequent actual and predicted tags overlap with those of Topic 2, including "Peace," "International Affairs," "Internationalism," and "War." The multilabel classifier seemed to pick up on some of the nuances connecting Topic 11 to the United States government and politics, assigning the tags "Politics" and "Government" to a number of test documents belonging to Topic 11. However, tags such as "Peace," "Internationalism," and "International Affairs" are still the top three tags that the classifier assigned to documents belonging to Topic 11. These tags belong to Topic 2. That means that more documents labeled with these tags belong to Topic 2 than to Topic 11. This is unsurprising since Topic 2 was the most frequently occurring topic in the dataset, while Topic 11 was the least frequently occurring topic. In other words, more documents had Topic 2 as their dominant topic than Topic 11. Therefore, there was a greater chance of a tag being assigned to Topic 2 since this assignment was based on the topic associated

with the highest percentage of documents labeled with that tag. In fact, the only tags that were assigned to Topic 11 were "Arbitration," "Courts," and "Mexico." This explains why there was a low degree of overlap between the predicted tags for documents belonging to Topic 11 and the tags associated with this topic.

In the case of Topic 11, the tags that this topic suggested should be used to label documents primarily made up of Topic 11 did not always align well with the tags the multilabel classifier suggested should be used. This was a result of how tags were assigned to topics; some topics might have claimed certain tags that were predicted often for documents belonging to different topics, as in the case of Topics 2 and 11. Therefore, assigning each tag to only the topic that occurred most frequently across documents with that tag might fail to capture some of the nuances in different topics that relate to similar themes. However, as stated earlier, the multilabel classifier appeared to capture some of the subtle differences between Topics 2 and 11, assigning tags such as "Government," "Politics," and "Legislation" frequently to documents belonging to Topic 11, but not to documents belonging to Topic 2. This again suggests that both the topic model and multilabel classifier convey similar kinds of information about the content of documents. Belonging to Topic 11 indicates that documents are mostly made up of content relating to the United States government, politics, and war since these are the themes that Topic 11's associated words convey. The multilabel classifier confirmed that documents belonging to Topic 11 contain this kind of content through the tags it predicted. Both the topic model and multilabel classifier reached similar conclusions through their analysis of the documents.

A closer examination of Topic 14 reinforces this idea that the multilabel classifier and topic model convey similar kinds of information about the texts. Topic 14 relates to children and child labor, as the words making up this topic suggest, shown in Figure 17. Many of the tags that

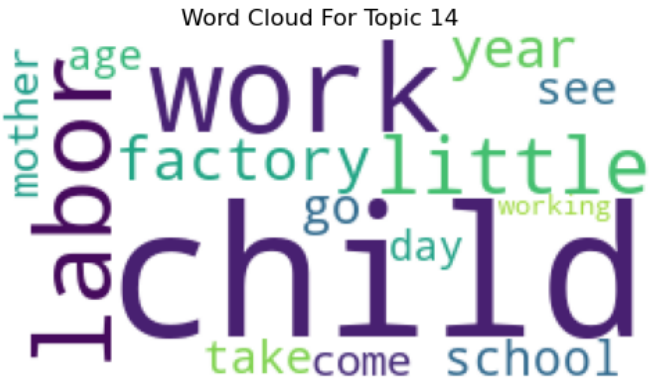
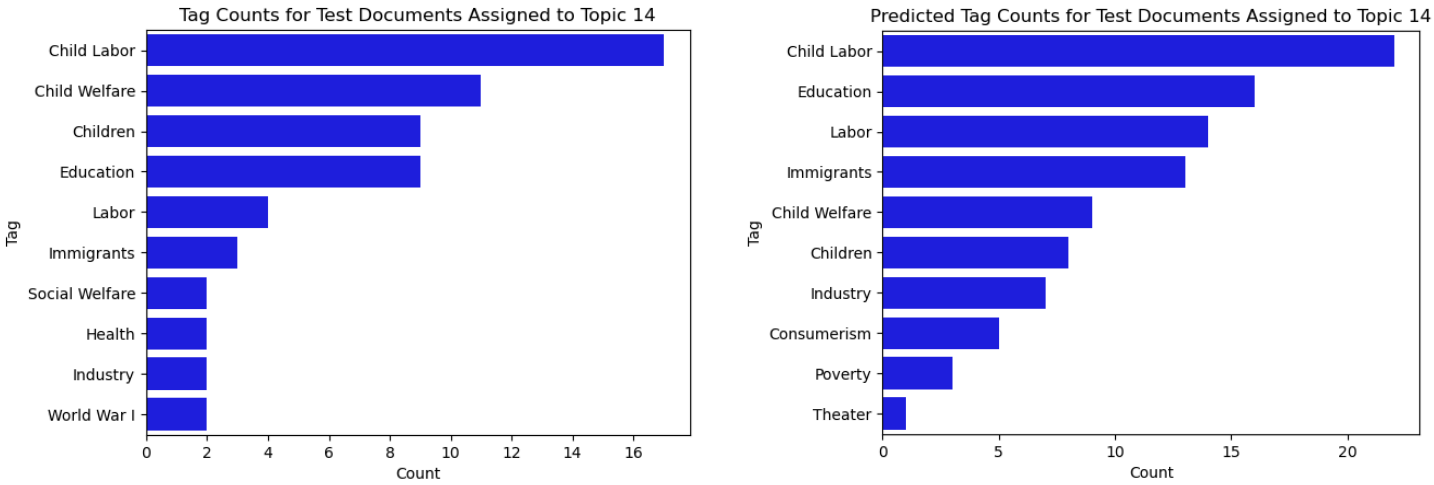


Figure 17. Words representing Topic 14, sized by associated word probabilities



(a) **(b)**

Figure 18. (a) Top 10 actual tag frequencies for test documents belonging to Topic 14 and (b) Top 10 predicted tag frequencies for test documents belonging to Topic 14

were both actually assigned and predicted to be assigned to documents in the test set belonging to Topic 14 relate to the theme of this topic. As shown in Figure 18 (a), tags such as "Child Labor," "Child Welfare," "Education," "Labor," and "Children" were frequently assigned to documents belonging to Topic 14. Figure 18 (b) indicates that the multilabel classifier predicted these tags often for these documents, just with different frequencies than what was actually the case for some tags.

Several of these frequently assigned tags, including “Child Welfare,” “Child Labor,” “Children,” and “Industry,” are also in the list of tags that were assigned to Topic 14. However, there are other tags that have been assigned to Topic 14 that did not appear frequently within the set of predicted tags for documents belonging to this topic, such as “Consumerism,” “Jobs,” “Publishing,” and “Manufacturing.” The case for Topic 14, then, is in between that of Topic 2, in which case all of the top tags predicted for documents belonging to Topic 2 had been assigned to that topic, and Topic 11, in which case none of the top tags predicted for documents belonging to Topic 11 had been assigned to that topic. The overlap between the predicted tags for documents belonging to Topic 14 and the tags assigned to Topic 14 reflects this middle position between Topics 2 and 11, with not quite as much overlap as what occurred for Topic 2, but more than what occurred for Topic 11. On average, about 45.53% of the predicted tags for documents belonging to Topic 14 also appeared in the set of tags assigned to Topic 14. Therefore, a little less than half of the time, the tags that Topic 14 suggested should be assigned to these documents were also the tags that the multilabel classifier suggested. There was also an association between the actual tags assigned to these documents in the Digital Edition and their dominant topic, with 52.52% of their actual tags also belonging to Topic 14. This further implies that a text’s dominant topic and the tags that both the multilabel classifier and a *Jane Addams Papers Project* staff member assigned to it convey similar information about its content.

Since there is a connection between the topics captured in documents and the labels assigned to them, it is possible that topic modeling and multilabel classification could be used together to summarize documents and organize them into different categories. If a classifier predicts tags for a document that belong to its dominant topic, those tags might best convey its content since both the classifier and topic model suggest that those tags should be used based on

the words contained in the document. The tags associated with a document's dominant topic could also aid individuals working on the *Jane Addams Papers Project* by providing an idea of what labels align with that document's content. The dominant topics of documents could also be used as an additional way to categorize documents, with the topics themselves being used as labels.

Since the dominant topics of documents align with the tags that both the classifier and *Project* staff members assigned to them, there are similarities between the topics that the topic model produced and the pool of tags that *Project* researchers have created over time. In other words, the themes that the topic model captured reflect the themes that *Project* researchers have identified in the documents and encapsulated in the tags. This means that the machine-generated insight about the content of Jane Addams' speeches and articles is similar to the insight generated by humans. However, since 42.70% of predicted tags and 45.10% of actual tags belonged to the same topic as their corresponding documents, about half of the time, a tag did not overlap with the document's dominant topic. This could mean that there is room for the creation of new tags that align with the themes found in the documents. Or, machine learning techniques might be unable to fully capture the meaning that humans can pull from texts. There are complexities and nuances contained in language that humans might be able to interpret more successfully than machines.

Machine-based methods like topic modeling and multilabel classification are still valuable, however, for textual analysis. Performing both of these techniques on Jane Addams' speeches and articles enabled the exploration of whether the topics and tags of documents were related. The multilabel classifier often predicted tags that appeared to relate to a document's dominant topic. This was captured better in the tag-topic overlap calculation for some topics

better than others as a result of how tags were assigned to topics. The connection between the themes that the topic model captured in the documents and the categories the multilabel classifier predicted for them suggests that these two methods can be used together to enhance an analysis of texts. Individually, topic modeling and multilabel classification contributed to the analysis of documents from the *Jane Addams Papers Project*. The LDA topic model uncovered 15 themes in the documents, including women's suffrage and international affairs. These themes captured the main ideas in the documents, while also highlighting more nuanced areas of meaning. Though the Binary Relevance Multinomial Naive Bayes classifier was not correct as often as desired, the model also offered an indication of the main ideas found in the documents through the tags that were predicted.

Conclusions

The goals of this research were to implement topic modeling and multilabel text classification on documents from Ramapo College's *Jane Addams Papers Project*, as well as to explore methods of bridging these two machine learning techniques when analyzing texts. The Digital Edition of the *Jane Addams Papers Project* contains a vast collection of documents relating to Jane Addams and her legacy as an activist and social worker. The other features of the Digital Edition, such as the tags assigned to documents and the collections of people, places, and organizations that connect to Jane Addams, provide a detailed, informative backdrop for the study of Addams and her influential role in history. The aim of using topic modeling and multilabel classification was to contribute to the wealth of information contained in the Digital Edition and investigate additional methods for the analysis and organization of its documents.

Through LDA topic modeling, Jane Addams' speeches and articles from the Digital Edition were analyzed simultaneously, allowing for the extraction of information from these texts and the discovery of patterns within them. The topic model with 15 topics identified themes within the documents, such as women's suffrage, international affairs, and child labor. These themes provide an understanding of the main ideas within the documents, serving as an overview of their content. Some of the topics were more distinctive, such as the one relating to women's suffrage. Other topics were connected to each other, but were differentiated by slight nuances in meaning. This suggests that there is overlap between the contents of documents; they may discuss similar subjects, but in different contexts. Assigning a topic to each document, the one estimated to make up most of its content, provided a method for using the topic model to analyze individual speeches and articles within the whole collection. The topics assigned to the

documents served as an additional source of information conveying their content that could be used alongside other descriptors of the documents, including the tags.

The multilabel classifier provided an automated method of assigning tags to documents. The Binary Relevance Multinomial Naive Bayes model that was trained on data split using the MultilabelStratifiedShuffleSplit class resulted in the best performance. On average, the model predicted about 32.87% of labels correctly for each document. A higher percentage of correct predictions would have been more desirable. Nevertheless, this multilabel classifier could serve as a tool to aid in the labeling of documents by providing suggested tags. The classifier offered a machine-based perspective on what tags were appropriate for each document. This perspective provided a new view of the contents of documents and how they might be categorized.

Using both topic modeling and multilabel text classification on Jane Addams' speeches and articles resulted in the opportunity to compare the kinds of information these two methods captured from the documents. Each speech and article had a set of corresponding tags pulled from the Digital Edition, a set of predicted tags assigned by the multilabel classifier, and a topic based on the results of the LDA topic model. These components associated with each document offered information about its content. In order to examine whether there was any overlap between these pieces of information, each tag was assigned to a topic. Linking the tags and topics in this way served as a bridge between the topic modeling and multilabel classification, facilitating an analysis of whether there was a relationship between a document's topic and tags. Around slightly less than half of the time, a document's actual and predicted tags belonged to the same topic as the document. This suggests a slight overlap between a document's topic and tags, with this overlap being more pronounced for certain topics that were assigned to larger numbers of documents.

Though multilabel classification and topic modeling are two different machine learning techniques–one supervised, the other unsupervised–they can produce results that provide similar information about the contents of texts. Both methods can be used to analyze and categorize documents, either according to the themes that a topic model indicates are prominent in that document or the labels that a classifier predicts. Based on the results of this project, the themes produced by a topic model often correlate with a supervised classifier's predictions. Used together, topic modeling and classification can complement each other, potentially producing a greater understanding of the subject matter of texts.

Future work could include further exploration of how topic modeling and text classification could be used together to analyze texts. Combining these two methods to produce information about the *Jane Addams Papers Project* documents could potentially reveal whether there are any additional tags that should be used to categorize the texts. It is possible that other labels that have never been used before may align well with the content of documents. Uncovering new ways to describe the documents could supplement the valuable, extensive collection of information contained in the *Jane Addams Papers Project* Digital Edition.

References

About Jane Addams. (n.d.). Jane Addams Papers Project. https://janeaddams.ramapo.edu/about-jane-addams/

About the project. (n.d.). Jane Addams Papers Project. https://janeaddams.ramapo.edu/about project/

Alamsyah, A., & Girawan, N. D. (2023). Improving clothing product quality and reducing waste based on consumer review using RoBERTa and BERTopic language model. *Big Data and Cognitive Computing, 7*(4), 168. https://doi.org/10.3390/bdcc7040168

Arslan, M., & Cruz, C. (2024). Business text classification with imbalanced data and moderately large label spaces for digital transformation. *Applied Network Science, 9*(1), 11. https://doi.org/10.1007/s41109-024-00623-5

Bird, S., Loper, E., & Klein, E. (2009). *Natural Language Processing with Python.* O'Reilly Media Inc. https://www.nltk.org/

Bradberry, T. J. (2018) *Iterative-stratification* (Version 0.1.9). https://github.com/trent-b/iterative-stratification

Decision Trees. (n.d.). Scikit-Learn. Retrieved March 18, 2025, from https://scikit-learn.org/stable/modules/tree.html#tree-classification

Detthamrong, U., Nguyen, L. T., Jaroenruen Y., Takhom, A., Chaichuay, V., Chotchantarakun K., & Chansanam W. (2024). Topic modeling analytics of digital economy research: Trends and insights. *Journal of Scientometric Research, 13*(2), 448–458. 10.5530/jscires.13.2.35

Dinsa, E. F., Das, M., & Abebe, T. U. (2024). A topic modeling approach for analyzing and categorizing electronic healthcare documents in Afaan Oromo without label

Information. *Scientific Reports (Nature Publisher Group), 14*(1), 32051. https://doi.org/10.1038/s41598-024-83743-3

Elghazel, H., Aussem, A., Gharroudi, O., & Saadaoui W. (2016). Ensemble multi-label text categorization based on rotation forest and latent semantic indexing. *Expert Systems with Applications*, *57*, 1-11. https://doi.org/10.1016/j.eswa.2016.03.041

Fujishiro, N., Otaki, Y., & Kawachi, S. (2023). Accuracy of the Sentence-BERT Semantic Search System for a Japanese Database of Closed Medical Malpractice Claims. *Applied Sciences, 13*(6), 4051. https://doi.org/10.3390/app13064051

An introduction to multilabel classification. (2020, July 16). GeeksforGeeks. Retrieved March 7, 2025, from https://www.geeksforgeeks.org/an-introduction-to-multilabel-classification/#

Jane Addams Digital Edition (n.d.). Jane Addams Papers Project. https://janeaddams.ramapo.edu/digital-edition/

Kholwal, R. (2023). Text-classify: A comprehensive comparative study of logistic regression, random forest, and knn models for enhanced text classification performance. *International Journal of Advances in Engineering & Technology, 16*(5), 415-433. https://doi.org/10.5281/zenodo.10148008

Lee, J., Chang, H. E., Cho, J., Yoo, S., & Hyeon, J. (2024). Analysis of issues related to nursing law: Examination of news articles using topic modeling. *PLoS One, 19*(8), 1–15. https://doi.org/10.1371/journal.pone.0308065

Li, J., Wang, S., Rudinac, S., & Osseyran, A. (2024). High-performance computing in healthcare: An automatic literature analysis perspective. *Journal of Big Data, 11*(1), 61. https://rdcu.be/ej66s

Luo L., & Li L. (2014) Defining and evaluating classification algorithm for high-dimensional

data based on latent topics. *PLoS One*, *9*(1): e82119. https://doi.org/10.1371/journal.pone.0082119

Mabey, B. (2023). *pyLDAvis* (Version 3.4.1). https://pypi.org/project/pyLDAvis/

models.ldamodel – latent Dirichlet allocation. (n.d.). Gensim. Retrieved March 15, 2025, from https://radimrehurek.com/gensim/models/ldamodel.html

Muthusami, R., Mani Kandan, N., Saritha, K., Narenthiran, B., Nagaprasad, N., & Ramaswamy, K. (2024). Investigating topic modeling techniques through evaluation of topics discovered in short texts data across diverse domains. *Scientific Reports (Nature Publisher Group), 14*(1), 12003. https://doi.org/10.1038/s41598-024-61738-4

Naive Bayes. (n.d.). Scikit-Learn. Retrieved March 18, 2025, from https://scikit-learn.org/stable/modules/naive_bayes.html#naive-bayes

Özmantar, M. F., Gökdağ, K., Hangül, T., & Agac, G. (2024). Research themes and trends in the field of teacher educators: A topic modelling study. *Teaching and Teacher Education*, *148*. https://doi.org/10.1016/j.tate.2024.104696

Pavlinek, M., & Podgorelec, V. (2017). Text classification method based on self-training and LDA topic models. *Expert Systems with Applications, 80*, 83–93. https://doi.org/10.1016/j.eswa.2017.03.020

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

Ploomber (2023). *Micro and macro averaging*. Sklearn-Evaluation. Retrieved March 18, 2025, from https://sklearn-evaluation.ploomber.io/en/latest/classification/micro_macro.html#

RandomForestClassifier. (n.d.). Scikit-Learn. Retrieved March 18, 2025, from https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

Rehurek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks* (pp. 45-50). ELRA.

Sechidis, K., Tsoumakas, G., & Vlahavas, I. (2011). On the stratification of multi-label data. *Lecture Notes in Computer Science*, *3*, 145-158. 10.1007/978-3-642-23808-6_10

Shaikh, R., Rafi, M., Naeem, A. M., Sulaiman, A., & Shaikh, A. (2023). A filter-based feature selection approach in multilabel classification. *Machine Learning : Science and Technology, 4*(4), 045018. https://doi.org/10.1088/2632-2153/ad035d

Tags. (n.d.). Jane Addams Papers Project Guide. http://jappg.wikidot.com/tags

Tandon, K., & Chatterjee, N. (2022). Multi-label text classification with an ensemble feature space. J*ournal of Intelligent & Fuzzy Systems, 42*(5), 4425–4436. https://doi.org/10.3233/JIFS-219232

Transforming the prediction target (y). (n.d.). Scikit-Learn. Retrieved March 17, 2025, from https://scikit-learn.org/stable/modules/preprocessing_targets.html#preprocessing-targets

Voskergian, D., Jayousi, R., & Yousef, M. (2024). Topic selection for text classification using ensemble topic modeling with grouping, scoring, and modeling approach. *Scientific Reports (Nature Publisher Group), 14*(1), 23516. https://doi.org/10.1038/s41598-024-74022-2

Wang, Q., Olshin, J., Vijay-Shanker, K., & Wu, C. H. (2023). Text mining of CHO bioprocess

bibliome: Topic modeling and document classification. *PLoS One, 17*(4), 1–12. https://doi.org/10.1371/journal.pone.0274042

Yuan, L., Xu, X., Sun, P., Yu, H. p., Yin, Z. W., & Zhou, J. j. (2024). Research of multi-label text classification based on label attention and correlation networks. *PLoS One, 19*(9). https://doi.org/10.1371/journal.pone.0311305

Zadgaonkar, A., & Agrawal, A. J. (2024). An approach for analyzing unstructured text data using topic modeling techniques for efficient information extraction. *New Generation Computing, 42*(1), 109–134. https://doi.org/10.1007/s00354-023-00230-5

Zhang, Y., Li, X., Liu, Y., Li, A., Yang, X., & Tang, X. (2023). A multilabel text classifier of cancer literature at the publication level: Methods study of medical text classification. *JMIR Medical Informatics, 11*. https://doi.org/10.2196/44892

Appendices

Topic Number	Associated Tags
0	'Sex Hygiene', 'Prisons', 'White Slavery', 'Illinois', 'Social Purity', 'Science', 'Legislation', 'Theater', 'Books', 'Civil Rights', 'Prostitution', 'Morality'
1	'Music', 'Public Works', 'Urban Planning', 'Sociology', 'Love', 'Business', 'Recreation', 'Education', 'Gambling', 'Illinois', 'Sweden'
2	'Germany', 'Disarmament', 'Women', 'Socialism', 'Hungary', 'World War I', 'Nobel Prize', 'Communism', 'Thanks', 'Economics', 'Internationalism', 'France', 'Military', 'Food Shortages', 'Peace', 'Holidays', 'Health', 'Russia', 'Agriculture', 'Europe', 'Love', 'Meetings', 'Austria', 'Canada', 'Conferences', 'International Affairs', 'Taxes', 'China', 'Revolution', 'Ethics', 'Relief Efforts', 'Pacifism', 'Soviet Union', 'Help!', 'Food Conservation', 'Public Opinion', 'Diplomacy', 'Serbia', 'Propaganda', 'Onsite', 'League of Nations', 'War', 'Foreign Policy', 'Economy', 'United States', 'Romania', 'Requests'
3	'Jobs', 'Home Economics', 'Ethnic Groups', 'Writing', 'Fashion', 'Hull-House Visits', 'Memberships'
4	'Film', 'Journalism', 'Microfilm', 'Gratitude', 'Settlement Movement', 'Hull-House Residents', 'Art', 'Health', 'Chicago', 'Introduction', 'Newspapers', 'Settlements', 'Employment', 'Poverty', 'Hull-House Visits', 'England', 'Housing', 'Family', 'Visits', 'Biography', 'Humor', 'Tributes', 'Neutrality', 'Architecture', 'Jobs', 'Medicine', 'Gossip', 'Hull-House', 'Fashion', 'Sanitation', 'Publications'
5	'Finance', 'Labor', 'Transportation', 'Health', 'Social Work', 'Unemployment', 'Public Health', 'Social Reform', 'Social Welfare', 'Philanthropy', 'Finances', 'Drugs'
6	'Abolition', 'Woman Suffrage', 'Politics', "'Women's Rights", 'Human Trafficking', 'Criticism', 'Gender Roles', 'Censorship', 'History', 'India',

	'Recommendations', 'Burma', 'Cartoons', 'Hawaii', 'Housing', 'Japan', 'Library', 'Prohibition', 'Eugenics', 'Research', 'African-Americans', 'Prisoners of War', 'Progressive', 'Racism', 'Temperance', 'Government', 'Social Justice'
7	'Charities', 'Feminism', 'Celebrations', 'Clubs', 'Awards and Honors', 'Legislation', 'Psychology', 'Tributes', 'Temperance', 'Poetry', 'Memberships', 'Civil Service', 'Charity'
8	'Abolition', 'Praise', 'Sociology', 'Ethnic Groups', 'Philippines', 'Refugees', 'South Africa', 'Visits', 'Disability', 'Ireland', 'Czechoslovakia', 'Books', 'Museums', 'Law', 'Finances', 'Sanitation', 'Turkey'
9	'Nationalism', 'Plays', 'Famine', 'Social Class', 'Crafts', 'Censorship', 'Immigrants', 'Anti-Semitism', 'Italy', 'Hull-House Visits', 'Free Speech', 'Anarchism', 'Eugenics', 'Immigration', 'Music', 'African-Americans', 'Greece', 'Prisoners of War', 'Anti-radicalism', 'Civil Rights'
10	'Sociology', 'Social Purity', 'Anti-Semitism', 'Patriotism', 'Love', 'Preparedness', 'Censorship'
11	'Mexico', 'Arbitration', 'Courts'
12	'Music', 'Race', 'Travels', 'Biblical figures', 'Friends', 'Eulogies', 'Historical figures', 'Visits', 'Literature', 'Philosophy', 'Religion', 'Democracy', 'Weddings', 'Poetry', 'Memberships', 'Illinois', 'Death'
13	'Crime Enforcement', 'Public Works', 'Sports', 'Transportation', 'Police', 'Juvenile Delinquency', 'Youth', 'Eugenics', 'Sanitation', 'Crime'
14	'Consumerism', 'Jobs', 'Publishing', 'Children', 'Child Labor', 'Manufacturing', 'Child Welfare', 'Industry'

Table 7. Tag assignments for each topic. Assigned according to the topic that is dominant for the greatest percentage of documents with a given tag