Leveraging Multiple Representations of Topic Models for Knowledge Discovery

COLIN M. POTTSM1, AKSHAT SAVALIYA1,2,3, AND ARNAV JHALA1, (Member, IEEE)

1Department of Computer Science, North Carolina State University, Raleigh, NC 27695, USA
2Khoury College of Computer Science, Northeastern University, Boston, MA 02115, USA
3Amazon Web Services (AWS), Dallas, TX 75240, USA

Corresponding author: Colin M. Potts (cmpotts@ncsu.edu)

This work was supported in part by the Laboratory for Analytic Sciences (LAS).

ABSTRACT Topic models are often useful in categorization of related documents in information retrieval and knowledge discovery systems, especially for large datasets. Interpreting the output of these models remains an ongoing challenge for the research community. The typical practice in the application of topic models is to tune the parameters of a chosen model for a target dataset and select the model with the best output based on a given metric. We present a novel perspective on topic analysis by presenting a process for combining output from multiple models with different theoretical underpinnings. We show that this results in our ability to tackle novel tasks such as semantic characterization of content that cannot be carried out by using single models. One example task is to characterize the differences between topics or documents in terms of their purpose and also importance with respect to the underlying output of the discovery algorithm. To show the potential benefit of leveraging multiple models we present an algorithm to map the term-space of Latent Dirichlet Allocation (LDA) to the neural document-embedding space of doc2vec. We also show that by utilizing both models in parallel and analyzing the resulting document distributions using the Normalized Pointwise Mutual Information (NPMI) metric we can gain insight into the purpose and importance of topics across models. This approach moves beyond topic identification to a richer characterization of the information and provides a better understanding of the complex relationships between these typically competing techniques.

INDEX TERMS Artificial intelligence, big data applications, data analysis, data visualization, knowledge discovery, natural language processing.

I. INTRODUCTION

In large unstructured and semi-structured text datasets, the process of knowledge discovery and information retrieval requires a way for analysts to identify content and categorize documents for tasks like summarization, anomaly detection, and report generation. Data filtering and triage requires analysts to downselect information from data streams to find relevant information. Analysts do not have a clear search target in this phase which results in an iterative process [1]. This is usually characterized by transitions between high-level exploration followed by a deeper look into elements of interest. As patterns are observed at a high-level, this leads to a deeper dive into specific filters along relevant dimensions. Finally, analysts then have to document and create summary reports for further discussion with other analysts and consumers. Often integration of information from segments that are assigned to multiple analysts is condensed into reports for final outcomes of the investigation [2], [3].

To address one of the challenges for analysts to extract semantically similar documents that lead to meaningful exploration, this paper proposes a process for analyzing topics between both the semantic features space of words and in the semantic feature space of documents. We demonstrate this approach on the Enron email dataset that contains over 500K organizational emails over several years. We design a
categorization task of identifying emails with informational content versus logistical content. No prior work known to the authors has addressed this specific task. To relate the feature spaces of words to documents, we utilize the Normalized Pointwise Mutual Information (NPMI) metric and show that it is effective in the categorization task. Overall, the main contributions of this paper are: (a) to introduce a novel document analysis task (section II.) (b) propose an initial approach that combines two types of topic models that have not previously been combined (section III.) and provide a rigorous theoretical justification of this approach (section IV.) (c) report insights gained from the model at a richer level than what is afforded by individual approaches (section V.) As part of evaluation and practical applicability, we worked extensively with a team from the U.S. National Security Agency’s Laboratory for Analytic Sciences comprised of two career analysts and two data scientists on a similar data repository that is unavailable to the public.

Topic discovery algorithms have been popular in characterizing the content in large text datasets. One of the most popular approaches is Latent Dirichlet Allocation (LDA) [4], which provides a set of distributions of related terms based on frequency of co-occurrence. These groups of highly correlated terms form topic groups that generally have semantic coherence. The algorithm is popular partly due to the interpretability of the resulting topic clusters and the controllability of features of the model. It is documented that parameter tuning is important for reproducible results [5] and the Dirichlet priors warrant more exploration [6]. LDA and similar probabilistic generative models label each document with a mixture of topics. These labels give an understanding of document relationships to topics and are of some use for comparing document content. Topics themselves are defined by groups of frequently co-occurring terms.

In addition to these term-based bag-of-words topic models, another approach for understanding documents is to use a low-dimensional neural embedding such as paragraph vectors [7]. Semantically similar documents then form clusters in the embedding space. These neural approaches are becoming more popular due to their ability to capture context of term-use and domain-specificity of terms.

Interpreting the output of these models remains an ongoing challenge for the research community. Consider figure 2 which show the topic distribution of a dataset versus two individuals of interest. This could be useful information for an analyst, but only if they have a semantic understanding on the underlying topics. This can be difficult to achieve generally and there are some specific drawbacks to the approaches models discussed here. For the LDA approach, due to the use of bag-of-words, the context in which a given term is used can only be extracted on inspection of specific documents. For instance, if we are considering a corpus of email messages then the topic word clusters indicate which emails include them but fail to capture the communicative intent. Second, it is unclear how much cross-topic interaction exists in the dataset within and across documents. For instance, in a dataset of emails, personal topics could be included in the text of the same email as work-related communication. For co-workers who have personal relationships, these same emails are either categorized in multiple topic categories or the topic groups contain terms from both types of interactions. Document clustering approaches, on the other hand, perform well if documents have few dominant topics. In cases where documents have multiple equally important topics, analysts need to do segmentation in order to derive insight about each topic from relevant document cluster.

We begin from the insight that these distinct models relying on different feature spaces are able to capture patterns in...
the underlying data from different functional perspectives. Beyond using the models individually however, an analyst needs to be able to understand the relationship between these models in order to accurately use them for retrieving information and interpreting results. In particular, when triaging large streams of email or other data a particular concern is a topic model’s ability to differentiate between the informational content of an email (the subject matter) and the logistical content or practical intent of an email (like scheduling, verifying presence of attachments, etc.). This type of categorization remains a challenge for each of these approaches individually.

**Dataset:** In support of combining topic models for analysts, we propose a method to map the abstract spaces of terms and documents. We first describe how we can characterizing the importance of topics and documents across the two spaces. We then use NPMI to characterize the relationship between topic clusters to identify different types of topics. We show the applicability of this approach through the analysis of over 500K emails from the Enron dataset [8]. In order to demonstrate the usefulness of discovering insight from the connection of these two abstract spaces, we provide results from the analysis of the Enron email dataset. The Enron email dataset [8] consists of > 500K email communications over 3.5 years capture from 150 accounts with nearly 30K internal and 77K total email addresses. These emails consist of around 18M words of which about 165K are unique words. The emails contain an average of 77 words. This dataset is well-utilized within computer science from several different perspectives including classical symbolic and modern neural approaches for language processing. There are a number of distinct projects within the area of Natural Language Processing such as topic discovery, sentiment analysis, stance analysis, etc. This dataset is also popular for social network analysis of organizational networks and communication patterns within these networks.

**II. TASK DESCRIPTION: LOGISTICAL VS INFORMATIONAL CONTENT**

For a clear definition of the task, we worked closely with a team of two analysts and two research scientists at the National Security Agency’s Laboratory for Analytic Sciences. Over a year of observations of their analysis process, we discovered that they were spending significant amount of time on triage of messages with informational content (such as arguments, question/answers, predictions, interpretations, general content-specific discussions, etc.) from logistical messages (such as trip planning, scheduling, acknowledgements, etc.). This inspired our task description and implementation on the Enron dataset. The intuition for this approach is to utilize the frequency and quantity of topical informational messages within structured organizational groups in the Enron dataset to improve the knowledge discovery process. This is achieved by extracting term clusters of common topics within organizational groups (functional teams) separately from topics that relate to more generic communication common throughout the organization (logistical content).

For topic modeling we use Latent Dirichlet Analysis (LDA) [4] over all the emails. This forms the basis of our topic model. The LDA algorithm provides clusters of related terms representing topics. For temporal data such as emails, overall topics are not of much interest to analysts. We further extend this model, following the work of TIARA [9], to compute topic strength over time. This allows analysts to filter emails based on topic across different time frames through an interactive interface. Figure 3 shows terms occurring in 5 topic groups. This indicates that topics 1 and 3 relate to relevant content (Topic 1: gas, market, oil, …; Topic 3: Enron, energy, news, …), whereas topics 2 (thanks, know, attached, …) and 4 (meeting, request, report, …) are related to logistical aspects or acknowledgements. Topic 5 is also informational content but corresponds primarily to a group of executives playing fantasy football. This provides a useful filter over the dataset for human analysts.

These topic clusters are currently analyzed with the assumption that they are independent. We are particularly interested in the question of the purpose of emails in terms of its relationship with topic clusters. Emails contain topical content, communicative acts (acknowledgement, approval, etc.), and logistics information.

**A. SIMILAR NLP TASKS**

While the authors believe this is a relatively novel task, it falls under the larger umbrella of text classification. Here we are particularly interested in unsupervised approaches to the task but consider a wide array of techniques in related work section. One prominent (and similar) task is spam detection. Many approaches to these more general tasks rely on supervised learning with large datasets of labeled data. This is particularly important for spam detection. Currently, there is no known dataset for the informational vs logistical content task as formulated here. Some conceptually similar ad-hoc tasks exist such as identifying questions or extracting calendar events information from emails, but we are looking at the problem more generally.

**III. ABSTRACT FEATURES SPACES**

**A. TERM SPACE: BAG-OF-WORDS TOPIC MODELS**

Bag-of-words document representations use a high-dimensional feature vector space with 1 dimension per unique word or term. Each feature has a value proportional to the count of that word appearing in a document. This could be a simple frequency count, or something more sophisticated like TD-IDF [10] where words are discounted based on their frequency in the entire collection. Each word is represented as a one-hot vector. Thus for $N$ unique terms in a model we get $N$ vectors $w_1, \ldots, w_N$ where $w_i$ has the $i$-th component set to 1 and all others are 0. This vector space is not used directly, but instead is the input to one or more topic modeling algorithms.

A popular algorithm using this representation is Latent Dirichlet Allocation [4]. This family of generative
probabilistic topic models are based on the assumption that the terms in a document are determined by a distribution that is dependent on the mixture of topics present in that document (see Figure 4). Each topic is assumed to be a distribution over terms in the corpus $\phi_k \sim \text{Dir}(\beta)$ where $\text{Dir}(\beta)$ is a Dirichlet distribution. Each document is modeled as being generated by choosing $\theta_i \sim \text{Dir}(\alpha)$ which is a mixture of topics $\theta_i$. The $j$-th document word is generated by choosing a topic $z_{i,j} \sim \text{Multinomial}(\theta_i)$ and then a word $w_{i,j} \sim \text{Multinomial}(\phi_{z_{i,j}})$ from that topic.

Documents are classified using the topic mixture as their representation. This mixture can then be used to measure similarity between documents and as a tool to understand the content of a document. Note that the final vector space is a low-dimensional embedding of the documents, but separate from the original word and document representations. Moreover, LDA does not consider vector operations such as addition. This can be a useful proxy for human topic similarity judgements under reasonable conditions [11].

Bag-of-words vectors are interpretable because each dimension represents a single term. They are limited because all dimensions are treated independently and thus relations between words are not captured. Topic mixtures from LDA are interpretable because they encode proportions of each topic present in a document. Individual topics are defined to be distributions over corpus terms and thus require users to be familiar with the algorithm to semantically interpret the outputs. One method for improving interpretability is to summarize a topic by looking at the top $N$ terms and/or thresholding probability $0 \leq p \leq 1$. Another is by weighting term probabilities based on keyword entropy across topics [9].

The space of documents can also be considered to be clustered by LDA. We have a mixture $\theta_i$ for each document that corresponds proportionally to the amount of content considered to be drawn from each topic. Thus this mixture can be used as a proxy for membership in a topic cluster. This can either be considered directly as a proportional membership or by using a minimum threshold $0 \leq t \leq 1$. Since $\theta_i$ comes from a Dirichlet distribution, this membership is very sparse so a single document is often only a member a relatively small set of topics compared to the total number of topics.

B. DOCUMENT SPACE: NEURAL DOCUMENT EMBEDDINGS

Neural embeddings, first proposed by Bengio et al. [12], seek to solve the problem of statistical language modeling (as opposed to the above methods focusing on topic). These models learn from large quantities of naturally occurring text data and predict sequences of words in sentences which is an inherently high-dimensional problem. Nevertheless, state-of-the-art techniques such as word2vec [13] have been successful in applications across various fields. Building on these language models, distributed document representations such as doc2vec [7] model the interaction between the context of a document and the use of language within it. These models embed both words and documents into a unified low-dimensional vector space (typically 100-300 dimensions) allowing meaningful vector operations such as cosine distance to represent similarity.

Doc2vec is the extension of word2vec model to unsupervised learning of continuous representation for larger text and documents. Word2vec is a continuous vector representation of words on large datasets based on a Neural Network Language Model (NNLM). The objective of the word2vec model is to provide the probability of word $w_{\text{target}}$ from the given context words $w_{\text{context}}$. Each word is represented by a vector that serves as a feature for the learning algorithm. As in Figure 5, the neural network takes the context words $w_{\text{context}}$
as input, averages or concatenates them, then predicts \( w_{\text{target}} \). The vector representations are then optimized to solve this task.

### Prediction

The power of the resulting representation lies in its ability to perform vector calculations on the embeddings. For instance, in a well-trained word2vec model \( \mathbf{v}_{\text{Madrid}} - \mathbf{v}_{\text{Spain}} + \mathbf{v}_{\text{France}} \) is closer to \( \mathbf{v}_{\text{Paris}} \) than to any other word vector. This works in the other direction as well as seen by vector addition where \( \mathbf{v}_{\text{Germany}} + \mathbf{v}_{\text{Capital}} \) should be close to \( \mathbf{v}_{\text{Berlin}} \).

For doc2vec, in addition to the context words \( \mathbf{w}_{\text{context}} \), the current document is also represented in the vector. This allows the model to characterize the changing nature of language on a per-document basis, instead of imposing a purely global model. Because of the structure of the algorithm, these document vectors are embedded in the same vector space as words. Thus vectors operations between document and between words and documents also carry semantic meaning. This allows operations to measure the similarity of a document to a word, or to see how adding/subtracting a word from a document vector might change its meaning. Where bag-of-words based approaches fail to preserve proximity information, neural models like word2vec and doc2vec preserve this information. Because of the nature of this semantic space, we can then use algorithms like k-means or OPTICS [14] to cluster documents and/or words.

### IV. COMBININGTERM AND DOCUMENT SPACES

Several researchers have attempted to unify bag-of-words and document embeddings to achieve interpretable vectors without losing proximity information and taking advantage of the rich semantics of low-dimensional embeddings of both words and documents [15], [16]. Instead of developing a new algorithm for topic modeling or document representation, we consider the question of using multiple techniques and provide a framework for their combined usage. Consider for instance the work of Bhatia et al. [17] who used neural embeddings to label topics. Their approach begins by training both word2vec and doc2vec models on Wikipedia. They then use article titles as labels. This gives them a labeled semantic space, somewhat missing from the direct application of the word2vec/doc2vec approaches. The algorithm works by comparing each keyword to each labelled vector and computing the average.

\[
rel_{d2v}(a, T) := \frac{1}{|T|} \sum_{v \in T} \cos \left( E^d_{d2v}(a), E^w_{d2v}(v) \right) \tag{1}
\]

\[
rel_{w2v}(a, T) := \frac{1}{|T|} \sum_{v \in T} \cos \left( E^w_{w2v}(a), E^w_{w2v}(v) \right) \tag{2}
\]

where \( a \) is a potential label, \( T \) is a set of topic keywords, \( E^d_{d2v}(\cdot) \) is a document vector in the doc2vec model, \( E^w_{d2v}(\cdot) \) a word vector, and \( E^w_{w2v}(\cdot) \) is a word vector in the word2vec model.

Expanding out the dot product equivalence of cosine in equation 1 and noting that the dot product is distributive with respect to vector addition and homogeneous under scaling to gives equation 4. Noting that the second vector is a unit vector and reapplying the definition of cosine gives us equation 5.

\[
rel_{d2v}(a, T) = \frac{1}{|T|} \sum_{v \in T} \frac{E^d_{d2v}(a) \cdot E^w_{d2v}(v)}{|E^d_{d2v}(a)| \cdot |E^w_{d2v}(v)|} = \frac{E^d_{d2v}(a)}{|E^d_{d2v}(a)|} \cdot \frac{1}{|T|} \sum_{v \in T} \frac{E^w_{d2v}(v)}{|E^w_{d2v}(v)|} \tag{3}
\]

\[
= \cos \left( E^d_{d2v}(a), \frac{1}{|T|} \sum_{v \in T} E^w_{d2v}(v) \right) \tag{4}
\]

\[
= \cos \left( E^d_{d2v}(a), \frac{1}{|T|} \sum_{v \in T} E^w_{d2v}(v) \right) \tag{5}
\]

We are now computing an average over keyword vectors and directly comparing the result to document vectors. This gives us equation 6 which is an embedding algorithm from keyword space to doc2vec space.

\[
\frac{1}{|T|} \sum_{v \in T} E^w_{d2v}(v) \tag{6}
\]

Similarly, we can derive an embedding into word2vec space which yields equation 7.

\[
\frac{1}{|T|} \sum_{v \in T} E^w_{w2v}(v) \tag{7}
\]

Drawing on results by Bhatia et al. [17] of computing semantic labels, we see that these are embeddings of topics represented as keywords into a neural document model such as word2vec or doc2vec. We can now utilize this computed embedding vector. The computed vectors are close via cosine similarity to Wikipedia articles describing similar topics. It is important to note however, that prior to producing final labels in their original approach they utilized PageRank and other optimizations such as support vector machines on human-labeled data. This was for differentiating between labels and not placing the embedding properly within the vector space.

\[\text{1This formulation is also an improvement on the number of computations needed to calculate this measure. In the original formulation, } O(n|T|) \text{ operations are needed where } n \text{ is the number of articles. Using equation 5 this is reduced to } O(n + |T|) \text{ [17].}\]
Moreover, they found that taking the average of the corresponding vectors from a topic’s documents did not produce good results and indicates that naive embedding of the average topic document vector isn’t sufficient.

We have already discussed that keywords sets are effective summarizations of LDA topics. Thus we have shown that using the embeddings in equations 6 & 7 we can create a useful embedding of an LDA topic into word2vec or doc2vec embedding space. This allows us to directly utilize those vector spaces and corresponding operations such as addition, subtraction, and cosine similarity to words, documents, and clusters in the neural embedding. Both models can then be utilized together to gain a better understanding of a target dataset. This mapping now allows us to present the results of document clustering in a more interpretable way by showing the terms associated with the document clustering approach.

V. METRIC-DRIVEN ANALYSIS

We have shown how to embed LDA topics into doc2vec space which enables us to directly compare results from the two models. We consider a general topic model to be a classification of a set of documents where a single document can be classified into multiple topics. For doc2vec, we test this with a simple k-means classifier. For LDA, we use the classification approach discussed before where a document is in the i-th topic if the i-th component of its topic mixture is above a threshold 0 ≤ t ≤ 1.

By treating topic models as classifications of the document space, we remove dependence on the models themselves. Comparison of models can now be carried out by comparing classifications provided by these models. We start by asking an information-theoretic question: What information does one model contain about the other? To answer this we use an established metric Normalized Pointwise Mutual Information (NPMI).

A. NORMALIZED POINTWISE MUTUAL INFORMATION

NPMI is an information theoretic measure that relies on probabilities of co-occurrence of two events x and y. When interpreted in this way, it attempts to quantify the information we gain about x by observing y, or the information gained about y when observing x since the measure is symmetric. It ranges from −1 for events that never co-occur, to 0 in the middle for events that are independent, to +1 on the other extreme for events that always co-occur. Pointwise Mutual Information (PMI) is computed as

\[ \text{pmi}(x; y) := \log \frac{p(x,y)}{p(x)p(y)} = \log \frac{p(x|y)}{p(x)} = \log \frac{p(y|x)}{p(y)} \]

(8)

where typically base 2 logarithms are used. Then NPMI is defined as

\[ \text{npmi}(x; y) := \frac{\text{pmi}(x; y)}{h(x)} \]

(9)

where \( h(x, y) \) is the joint self-information of x and y.

\[ h(x, y) := - \log p(x, y) \]

(10)

B. COMPARING LDA AND DOC2VEC WITH NPMI

Instead of using sliding windows of words like the \( C_r \) measure, we are interested in measuring the NPMI of a document being classified in topic i in model A and topic j in model B. NPMI is calculated in our model as −1 for events that never co-occur, 0 for totally uncorrelated events, and 1 for events that always co-occur. When applied for a topic model against itself, we learn information about the correlation between topics. Consider the 10 topic model in Table 1. This model is from LDA trained on about 250k messages from the Enron collection. These messages contain on average about 77 words. From looking at the keywords (the top 8 by probability from the LDA model) we can readily see that topics 2 and 8 seem similar. As mentioned in figure 3, a model with more topics for the Enron dataset might make sense for use by an analyst. We have chosen 10 topics for LDA (and later 5 for doc2vec) for purposes of exposition only. In Figure 6 we note that the NPMI of those topics is .54. This means that topics 2 and 8 often co-occur in document topic mixtures. This is not a measure of similarity (for similarity we could compare distributions directly) but instead a measure of the relationship between topics in terms of how and when they are discussed. This succinct visualization is an example of how using NPMI for comparing topics gives a potential analyst an additional tool for understanding the relationships of topics in a model.

Looking further at the keywords in Table 1 and the correlations in Figure 6 it is difficult to make sense of why things co-occur and also difficult to determine which topics correspond with logistical content (such as scheduling, follow-ups, requesting attachments, etc.) and which are about topics with substance which we call content. A first step at making these more understandable for a human reader is to use another keyword selection algorithm such as the one from TIARA [9] which discounts keywords based on entropy across documents. This alternate view is presented...
TABLE 1. Keywords for 10 LDA topics trained on the Enron corpus of emails. These keywords are computed by taking the top 8 keywords based on the probability distribution of each LDA topic ($\phi$ in Figure 4).

| Topic 1       | Topic 2       | Topic 3       | Topic 4       | Topic 5       |
|---------------|---------------|---------------|---------------|---------------|
| enron         | new           | enron         | gas           | enron         |
| new           | company       | know          | way           | enron         |
| deal          | said          | thanks        | meeting       | energy        |
| thanks        | million       | let           | know          | trading       |
| need          | enron         | need          | thanks        | agreement     |
| gas           | click         | going         | information   | credit        |
| know          | free          | think         | new           | market        |
| time          | business      | pm            | list          | price         |

Table 2. Looking at this new view of our LDA topics, we note that it seems that this topic model does not distinguish between these types of messages. Namely, our topics all mix logistical with informational content. This makes the model more difficult to interpret and motivates our search for a better way mining this information from these messages.

We have discussed how a neural document embedding provides a rich semantic vector space to explore documents. For the Enron corpus we used a model trained on English Wikipedia to calculate document embeddings for each message. We then used a $k$-means classification of those vectors to classify messages into 5 disjoint clusters [18]. The keywords for these 5 topics are found in Table 3. These keywords were extracted using a modified version of the TIARA keyword extraction algorithm applied to the text of the clustered messages. This model is not a particularly advanced topic model [15], [16], [19]. We will show however, that it provides useful additional information when paired with an LDA topic model. In future applications, more advanced models could be substituted for each model.

The doc2vec model in Table 3 contains topic 2 which is a topic about logistical not informational content. This can be seen from keywords such as “thanks,” “meeting,” and “attached.” This model, unlike the LDA model in tables 1 and 2, denotes a cluster of messages about logistical content.

Returning to our analysis and we calculate NPMI scores across models to gain insight into the relationship between them. This information is presented in Figure 7. In particular, we see that we have two doc2vec topics interspersed throughout our LDA topics. The first, topic 2, as we have already seen is about logistics. Topic 5 consists almost exclusively error messages from a faulty tool used by some Enron employees, with some discussion about the tools but also chunks of actual emails present when the tool failed. This shows however, that by utilizing multiple distinct topic models and comparing their scores via NPMI, we can gain useful insight into each individual model and the target corpus as a whole. We have shown that the system identifies not only topics but also the purpose of emails by distinguishing logistical from informational content. This illustrates the ability of this approach to provide deeper insight to analysts than either existing approach individually. Distinguishing content from intent was a particular goal of our work and we have demonstrated that the novel approach provides deeper insight on this problem.

VI. RELATED WORK

Over the years, several approaches have been introduced for topic discovery on text data. Some of them are generative probabilistic topic models, vector representational models, and neural variational inference models. Miao et al. [20] presented the idea of using Neural Variational Inference to uncover the topics in text based documents. Their neural variational document model combines a continuous stochastic document representation with a bag-of-words generative model. When probabilistic generative model structure becomes deeper and more complex true Bayesian inference becomes intractable. The general idea of the paper is to build the inference network to approximate intractable distribution over the latent variables. Towne et al. consider how
TABLE 2. Weighted keywords for LDA topics on the Enron corpus. The weights for keywords \( w_j \) in topic \( k \) are calculated as 
\[
p(w_j|\phi_k) \times \left( \log p(w_j|\phi_k) - \frac{1}{K} \sum_{k=1}^{K} \log p(w_j|\phi_k) \right)
\]
according to the algorithm developed in TIARA [9]. We can see in these keywords, compared to those in Table 1, that topics that appeared to be about logistical content might not be. Thus an algorithm is needed to distinguish these messages and topics from those about actual informational content.

| Topic 1       | Topic 2       | Topic 3     | Topic 4     | Topic 5     |
|---------------|---------------|-------------|-------------|-------------|
| deal          | million       | detected    | way         | gas         |
| deals         | company       | enron       | gas         | trading     |
| request       | said          | know        | folder      | credit      |
| yahoo         | technology    | variances   | meeting     | financial   |
| sara          | round         | let         | roundtrip   | oil         |
| new           | investors     | going       | ectr        | price       |
| gas           | click         | key         | fares       | market      |
| new           | thanks        | tx          | capacity    |             |

| Topic 6 | Topic 7 | Topic 8 | Topic 9 | Topic 10 |
|---------|---------|---------|---------|----------|
| updated | power   |         | pm      | feedback |
| wr      | energy  |         | houston |          |
| rb      | california |       | parsing |          |
| fantasy | state    |         | million |          |
| qb      | electricity |      | court   |          |
| injury   | market   |         | billion  |          |
| agent   |         |         | stock   |          |
| week    |         |         | ancillary |         |

TABLE 3. TIARA-like keywords for doc2vec topics. Topic 1 are largely spam or newsletter messages without any real content. Topic 2 is about email communication and scheduling, which is what we call a topic about logistical content. Topic 3 is about fantasy football (several Enron executives emailed about it frequently). Topic 4 is about Enron’s main business (e.g., energy, regulations) and finally topic 5 is messages from a schedule crawler that was throwing errors with the snippet “error occurred while attempting to initialize the Borland Database Engine” repeated many times.

| Topic 1        | Topic 2       | Topic 3    | Topic 4     | Topic 5     |
|----------------|---------------|------------|-------------|-------------|
| click          | thanks        | fantasy    | enron       | 0.314      |
| unsubscribe    | meeting       | wr         | energy      | 0.060      |
| free           | attached      | game       | 0.08        |             |
| fares          | know          | updated    | power       | 0.037      |
| email          | let           | rb         | state       | 0.035      |
| information    | vince         | know       | electricity | 0.035      |
| travel         | initialize    | yards      | california | 0.033      |
| online         | bolrand       | qb         | company     | 0.024      |

human perception matches with LDA in terms of document similarity [11]. Zhang and Li [18] consider the general applicability of \( K \)-means clustering as applied to vector spaces and topic modeling, finding that it is a viable technique for inferring topics and helping to justify our choice of \( K \)-means as one of models of interest.

Moody et al. [16] introduced mixing Dirichlet-like models with word embeddings to create the lda2vec model. They used dense vectors to represent the semantic and syntactic regularities in the document and Dirichlet distributed latent document level mixture of topics to get sparse document representation with dense topic vectors. They modified the Skip-gram Negative-Sampling to utilize document wise featured vectors while simultaneously updating document weights onto topic vectors. Similarly, the Deep Word-Topic LDA trains an LDA [21] model but uses embeddings from a Skip-gram model instead of one-hot word representations. These represent a subset of techniques for combining embeddings with topic models, which contrast to our approach of comparing results of an ensemble of models. What these works share in common is a recognition that semantic interpretations of topics remain a difficult task and active area of research. The bag-of-concepts model addresses combining these models by clustering word vectors [15]. One key limitation in these previous approaches is that documents containing multiple topics are not correctly segmented although this is also an active field of research [22], [23], [24]. Another aspect is the loss of interpretability in converting LDA based approaches to word-embeddings. One approach to solve this limitation is a visual exploration of neural document embedding space [25]. Neural models also require significant amounts of data and parameter tuning [26]. In this work, we are relating the two approaches in a way that we can map the sparse vectors and dense vectors using NPMI metric to discover and segment different dominant topics in a given document. Other approaches have considered combining topic models such Schnober and Gurevych [27] who looked at using multiple LDA models for digital humanities research. Robertson et al. [2] used pointer generator networks [28] augmented with LDA models to segment topics in artificially mixed documents. One of the reason to do that is, generally documents consist of multiple topics and using this way we can segment document topics in a better way.

Instead of combining multiple topic models, Sterling & Montemore [29] combined citation networks and text similarity for to create a recommender system for research articles. Many topic modeling approaches are concerned with accurately modelling the underlying distribution of topics,
but our primary concern is the utility of these topics to a potential analyst. Korenčič et al. [30] also consider this viewpoint and specifically ask how model-induced topics correspond to topics of interest to an analyst. They compute a topic-coverage metric using a reference set of known-interesting topics. This contrasts to our approach of using an unsupervised algorithm to generate comparisons between multiple topic models, although we do impose a binary classification of topics related to semantic vs logistical content. Churchill and Singh [31] consider topic models from the perspective of topic-noise, using pre-trained model in an ensemble with LDA to generate more diverse and coherent topics. This approach also uses the semantic information from word embeddings but is specifically tuned to use topic-noise as a discriminator that can be used either during or after topic generation. Their work was primarily focused on short-form social media text, which was also explored by Zhang et al. [32] who combine FastText [33] embeddings with prior work on Sentence-LDA [34].

VII. DISCUSSION

This work proposes a general methodology for leveraging multiple document representation frameworks for analysts interested in triage of large semi-structured datasets. To that end, we have specifically looked at two popular paradigms, doc2vec and LDA. We have explicitly shown a direct mapping from LDA topics to doc2vec vector space (section IV). This is a novel application of a technique previously used only for topic-labeling of LDA topics. Prior work [17] indicates this mapping is semantically meaningful, and we show it can provide additional insight as a way of comparing semantics across these two different techniques. The primary validation of our work is on our proposed usage of NPMI to compare topics across models, agnostic to the underlying theoretical framework, by considering topic models to be clusters of documents. This is a distinct usage of NPMI from the \( C_v \) topic coherence metric. The NPMI scores from our test dataset show that we can use this combination of models to effectively discern information from logistical content. In particular, the visual paradigm used indicates which topics correspond with logistical content, something that neither model individually was able to do. It is possible for an analyst to make this discernment, as we showed when using top-k topic terms to help understand the NPMI output.

VIII. CONCLUSION

We have presented a methodology to relate abstract feature spaces of bag-of-words and document-embedding approaches. We have begun to explore the rich space of possible embeddings through our approach of directly embedding LDA topics into low-dimensional embeddings, but this work could benefit from a richer mathematical and empirical analysis of competing embedding techniques and their implications for the image of an LDA topic in the embedding space. In particular, one could view the lda2vec algorithm [16] as an alternative embedding approach that instead focuses on finding a set of basis vectors that generate all document vectors under a sparsity constraint along basis directions. Through a concrete demonstration on a well-studied dataset of emails from the Enron corpus, we show that the application of this methodology provides deeper insight into the discovery of knowledge about multiple topics in terms of their purpose (e.g., informational vs. logistical content) and in terms of inter-topic relationships. This provides a strong argument that analysts in information retrieval and knowledge discovery tasks should use systems specifically developed to leverage multiple topic models with different theoretical underpinnings in order to maximize the knowledge they gain during their investigations. We believe particularly that interactive analysis systems should feature this approach when allowing users to explore the space of potential topics models and when interpreting topics and topic distributions.

IX. FUTURE WORK

Future work for this topic includes a more rigorous exploration of the informational versus logistical information task. A strong baseline dataset for researchers to use when developing and evaluating new approaches would greatly benefit the community. Similar to how common datasets like the Twenty Newsgroups datasets and coherence metrics. Another area is the applicability of NPMI cross-model comparisons to other theoretical frameworks. We shows that this technique helps with this task using doc2vec and LDA, but is is possible that other cross-comparisons would yield insight into other related semantic questions.

ACKNOWLEDGMENT

Any opinions, findings, conclusion, or recommendations expressed in this article are those of the author(s) and do not necessarily reflect the views of the LAS and/or any agency or entity of the United States Government.

REFERENCES

[1] K. Wongsuphasawat, Z. Qu, D. Moritz, R. Chang, F. Ouk, A. Anand, J. MacInlay, B. Howe, and J. Heer, “Voyager 2: Augmenting visual analysis with partial view specifications,” in Proc. Conf. Hum. Factors Comput. Syst., May 2017, pp. 2648–2659.

[2] J. Robertson, B. Harrison, and A. Jhala, “Interactive summarization for data filtering and triage,” in Proc. 33rd Int. Florida Artif. Intell. Res. Soc. (FLAIRS) Conf., 2020, pp. 1–6. [Online]. Available: https://aaai.org/ocs/index.php/FLAIRS/FLAIRS20/paper/view/18440/17595

[3] C. M. Potts and A. Jhala, “Narrapot: Narrative-based interactions and report generation with large datasets,” in Interactive Storytelling, A. Mitchell and M. Vosmeer, Eds. Cham, Switzerland: Springer, 2021, pp. 118–127.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet allocation,” J. Mach. Learn. Res., vol. 3, pp. 993–1022, Jan. 2003.

[5] A. Agrawal, W. Fu, and T. Menzies, “What is wrong with topic modeling? And how to fix it using search-based software engineering,” Inf. Softw. Technol., vol. 98, pp. 74–88, Jun. 2018.

[6] H. Wallach, D. Mimno, and A. McCallum, “Rethinking LDA: Why priors matter,” in Proc. Adv. Neural Inf. Process. Syst., vol. 22, 2009, pp. 1–9.

[7] Q. Le and T. Mikolov, “Distributed representations of sentences and documents,” in Proc. Int. Conf. Mach. Learn., 2014, pp. 1188–1196.

[8] B. Klimt and Y. Yang, “The Enron corpus: A new dataset for email classification research,” in Proc. Eur. Conf. Mach. Learn. Cham, Switzerland: Springer, 2004, pp. 217–226.
F. Zhang, W. Gao, Y. Fang, and B. Zhang, “Enhancing short text topic modeling with FastText embeddings,” in Proc. Workshop Int. Conf. Learn. Represent. Learn. (ICBAIE), Jun. 2020, pp. 255–259.

P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, “Enriching word vectors with subword information,” Trans. Assoc. Comput. Linguistics, vol. 5, pp. 135–146, Dec. 2017.

F. Zhang, W. Gao, and Y. Fang, “News title classification based on sentence-LDA model and word embedding,” in Proc. Int. Conf. Mach. Learn., Big Data Bus. Intell. (MLBDBI), Nov. 2019, pp. 237–240.

COLIN M. POTTS received the B.A. degree with a double major in both mathematics and computer science from Lawrence University, Appleton, WI, USA, in 2013, and the M.S. degree in computer science from North Carolina State University, in 2019, where he is currently pursuing the Ph.D. degree in computer science.

In 2013 he was an Assistant Researcher at Smart Information Flow Technologies (SIFT) working on automated planning and real-time control systems. From 2014 to 2016, he worked at the Liquid Narrative Group at North Carolina State University as a Research Assistant focusing on multimedia narrative discourse planning. In 2017, he was at the Laboratory for Analytic Sciences on natural language processing and multimedia reporting for semi-structured text datasets. Since 2020, he has been with Dr. Jhala on interactive natural language processing systems for knowledge discovery and multimedia narrative generation.

AKSHAT SAVALIYA received the B.Eng. degree in computer engineering from Gujarat Technological University, India, in 2019, and the M.S. degree in computer science from Northeastern University, in 2022.

He has held several research positions including working on data analytics at Ahmedabad University, India, where he helped build a system for tracking market data of commodities in support of local farmers. He then worked on applications of topic modeling techniques at North Carolina State University, in 2020. In 2021, he joined as a Computational Science Intern at the Jackson Laboratory. In 2022, he joined as a Software Development Engineer with Amazon.

ARNAV JHALA (Member, IEEE) received the B.Eng. degree in computer engineering from Gujarat Technological University, India, and the M.S. and Ph.D. degrees in computer science from the North Carolina State University, in 2004 and 2009, respectively. He held academic positions at UC Santa Cruz, the IT University of Copenhagen, the Institute for Creative Technologies (ICT) at University of Southern California, and the Indian Space Research Organization (ISRO). He is an Associate Professor of computer science with the North Carolina State University and the Director of the Visual Narrative Interdisciplinary Faculty Cluster. His research interests include intersection of artificial intelligence and digital media, particularly in the areas of computer games, cinematic communication, and narrative discourse.

* * *