Information Processing and Management 57 (2020) 102216

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Learning entity-centric document representations using an entity facet topic model

Chuan Wu[*,a,b], Evangelos Kanoulas[b,c], Maarten de Rijke[b]

[a] School of Information Management, Wuhan University, Wuhan, Hubei, China
[b] Informatics Institute, University of Amsterdam, Amsterdam, The Netherlands
[c] Amsterdam Business School, University of Amsterdam, Amsterdam, The Netherlands

ARTICLE INFO

Keywords:
Document representation
Topic models
Entity aspects
Text classification

ABSTRACT

Learning semantic representations of documents is essential for various downstream applications, including text classification and information retrieval. Entities, as important sources of information, have been playing a crucial role in assisting latent representations of documents. In this work, we hypothesize that entities are not monolithic concepts; instead they have multiple aspects, and different documents may be discussing different aspects of a given entity. Given that, we argue that from an entity-centric point of view, a document related to multiple entities shall be (a) represented differently for different entities (multiple entity-centric representations), and (b) each entity-centric representation should reflect the specific aspects of the entity discussed in the document.

In this work, we devise the following research questions: (1) Can we confirm that entities have multiple aspects, with different aspects reflected in different documents, (2) can we learn a representation of entity aspects from a collection of documents, and a representation of document based on the multiple entities and their aspects as reflected in the documents, (3) does this novel representation improves algorithm performance in downstream applications, and (4) what is a reasonable number of aspects per entity? To answer these questions we model each entity using multiple aspects (entity facets[1]), where each entity facet is represented as a mixture of latent topics. Then, given a document associated with multiple entities, we assume multiple entity-centric representations, where each entity-centric representation is a mixture of entity facets for each entity. Finally, a novel graphical model, the Entity Facet Topic Model (EFTM), is proposed in order to learn entity-centric document representations, entity facets, and latent topics.

Through experimentation we confirm that (1) entities are multi-faceted concepts which we can model and learn, (2) a multi-faceted entity-centric modeling of documents can lead to effective representations, which (3) can have an impact in downstream application, and (4) considering a small number of facets is effective enough. In particular, we visualize entity facets within a set of documents, and demonstrate that indeed different sets of documents reflect different facets of entities. Further, we demonstrate that the proposed entity facet topic model generates better document representations in terms of perplexity, compared to state-of-the-art document representation methods. Moreover, we show that the proposed model outperforms baseline methods in the application of multi-label classification. Finally, we study the impact of EFTM's parameters and find that a small number of facets better captures entity specific topics, which confirms the intuition that on average an entity has a small number of facets reflected in documents.

* Corresponding author.
E-mail addresses: wu.chuan@whu.edu.cn (C. Wu), e.kanoulas@uva.nl (E. Kanoulas), derijke@uva.nl (M.d. Rijke).

[1] To avoid unnecessary ambiguity, we use facet instead of both aspect and facet across this work.

https://doi.org/10.1016/j.ipm.2020.102216
Received 30 September 2018; Received in revised form 23 January 2020; Accepted 26 January 2020

1. Introduction

Understanding the content of documents by learning semantic representations can benefit various downstream applications, such as information retrieval (Croft, 1981) and text classification (Sebastiani, 2002). Existing document representation methods include (1) traditional bag of words (BoW), which represent documents using term frequencies; (2) topic distributions (Blei, Ng, & Jordan, 2003), which represent documents using mixed distributions of latent topics; and (3) dense vector representations (Le & Mikolov, 2014), which represent documents as points in a low dimensional space.

Existing document representation methods assume a representation in terms of the semantic topics discussed in the document. However, when it comes to understanding a document from the perspective of entities, it is natural to think of a document in terms of the facets of the different entities it relates to. In other words, in this work we hypothesize that entities are not monolithic concepts; instead they have multiple facets, and different documents may be discussing different facets of a given entity. Given that, we argue that from an entity-centric point of view, a document related to multiple entities shall be (a) represented differently for different entities (multiple entity-centric representations), and (b) each entity-centric representation should reflect the specific facets of the entity discussed in the document. Let us illustrate this hypothesis with an example. Political world leaders, as entities, may have different facets of them described in news articles, such as family, foreign policy, campaigning, economy, etc. A document related to political world leaders is expected to reflect only specific facets of these leaders. For example, a document discussing a meeting between the presidents of the USA and Russia in 2016 might mention multiple entities, such as Obama and Putin. Economic and foreign policy facet of these entities are probably reflected in the document, but it is unlikely that presidential campaign facet or family facet are discussed.

The main objective of our work is to explore representing documents based on entity facets. In particular, to accurately represent a document, we propose to use entity specific topics that reflect the facet of the entity discussed in a document, which we call *entity facets*. Further, we define an *entity-centric document representation* as a distribution over entity facets. Each entity facet is further defined as a distribution over latent topics. Continuing our world leader example, Fig. 1 shows an example of an entity-centric document representation involving the two leaders. Document d_1 is associated with two entities, while d_0 and d_2 are each associated with a single entity. The entity-centric representation of each document is shown within the dashed rectangles above the documents; for every associated entity, a distribution over its facets is learned for that entity. As we can see, facet 1 of Barack Obama (BO) and facet 0 of Vladimir Putin (VP) are reflected in d_1, while for d_0 and d_2, the entity-centric representation also indicates the facet reflected for the corresponding source entity, i.e., VP and BO, respectively.

We propose a new task, that of entity-centric document representation learning. For a document associated with multiple entities, multiple facet distributions are learned for the document. To understand our modeling decisions and the contribution that we make, we list different topic models, with their prime ingredients: words, entities, topics, and facets in Fig. 2. Previously proposed topic models have considered words and topics (e.g., LDA, Fig. 2(a)); words and entities-as-sources-of-information (the Author Model, Fig. 2(b)); words, entities-as-observables and topics (Link-LDA, Fig. 2(c)); and words, topics, and entities-as-sources (the Author Topic

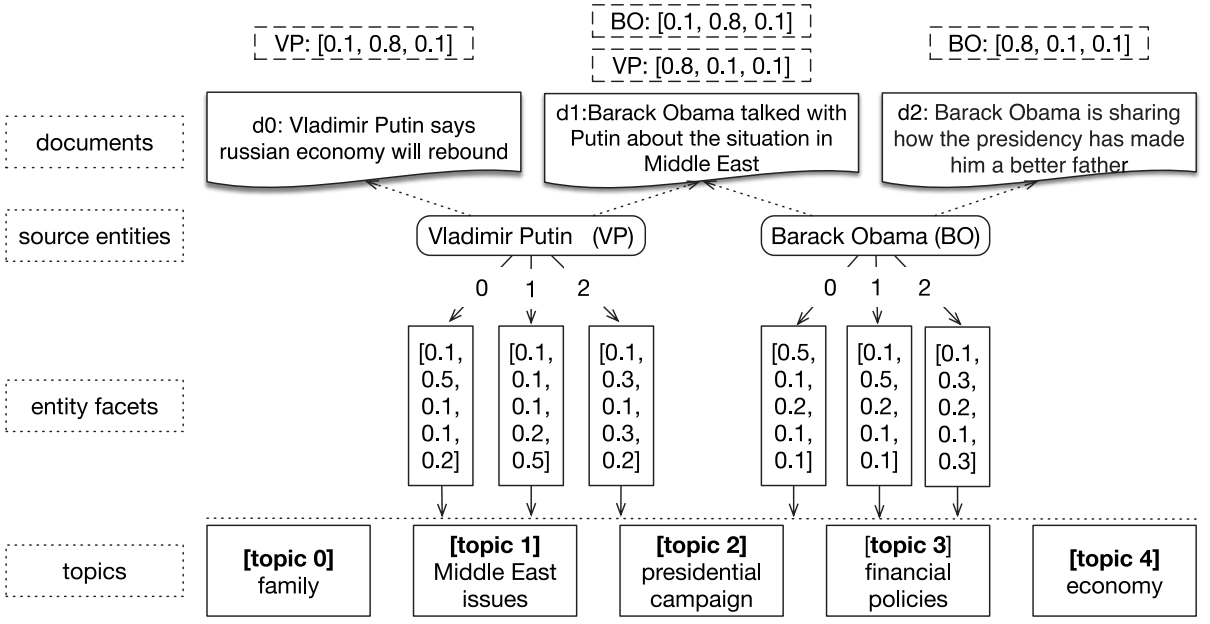

Fig. 1. Illustration of an entity-centric document representation. Documents d_0, d_1 and d_2 are associated with the sets of source entities {VP}, {BO, VP}, and {BO}, respectively. Each source entity has three facets. Each entity facet is a distribution over five latent topics, where each element of the distribution indicates the relatedness between the facet and corresponding topic. The entity-centric document representations are mixed proportions of facets shown above the corresponding documents.

Model, Fig. 2(d)). We propose the *Entity Facet Topic Model* (EFTM) (Fig. 2(e)) as a model for learning entity-centric document representations. In a generative perspective on representation learning, we model entities as a source of information (*source layer*); such source entities are assumed to be given and could be labels or metadata assigned to a document, or in-text entities; they consist of multiple entity facets (*facet layer*). Each facet is built upon general latent topics (*topic layer*), which are further used to generate observed variables (*observation layer*) in documents. Since different types of observed variables (e.g., words and entities) might appear in documents (Erosheva, Fienberg, & Lafferty, 2004) (Fig. 2(c)), we consider both words and entities. To differentiate between entities in the source layer and the observation layer, we refer to the former as source entities, and the latter as document entities.

To illustrate how our work can be used in practical applications, we present the following examples.

Example 1. Knowledge Base Construction/Population Fetahu, Markert, and Anand (2015) propose to populate Wikipedia entity pages by automated news article suggestion. Their approach consists of two steps, identifying articles relevant to entities and then connect relevant articles to particular sections. By learning entity-centric document representations using news articles labeled by their related entities (source entities), one might be able to cluster news articles into smaller groups on the per-entity basis. If groups of articles match contents of particular sections of a Wikipedia page, we might update the section. If it does not match, we might update the Wikipedia page by adding new sections using groups of articles identified as references.

Example 2. Entity Salience Detection Dunietz and Gillick (2014) propose a new entity salience task called entity salience detection, which aim at identifying whether an entity is salient in a document it appears in. If a particular facet of an entity is reflected in a document, it is likely to act as an important role in the document, and thus being considered as salient for the document. By learning entity-centric document representation of the document using a model trained by available datasets, we can perform binary classification from the perspective of the entity and tell whether the document matches any facet of the entity.

Example 3: Personalized News and Blog Recommendation Kazai, Yusof, and Clarke (2016) presents a prototype mobile app that provides personalized content recommendations to its users by combining various user signals. In the application, individual models are built for each user. Similarly, our model learns individual facets for each user from news and blogs that each user is interested at. Then, entity(user)-centric representations of documents are inferred and judged whether it should be recommended to each user.

Summarizing, EFTM models entities as sources of information along with their multi-faceted properties. At the same time, EFTM represents documents on the basis of the facet distributions of the source entities, which provides the desired entity-centric representations for documents. Our work aims at facilitating information processing applications, in which entities are central towards understanding, and modeling text. Examples include multi-authored article collections, multi-labeled textual collections and so on.

The main contributions of this work are as follows:

1. We propose the task of entity-centric document representation learning.
2. We propose a novel Entity Facet Topic Model (EFTM) to learn entity-centric document representations.
3. We confirm our hypothesis regarding the existence of multiple facets of an entity by analyzing the learned entity facets using qualitative and quantitative analysis, and identify a effective number of facets per entity.
4. We demonstrate the effectiveness of EFTM in downstream applications using a multi-label classification task.

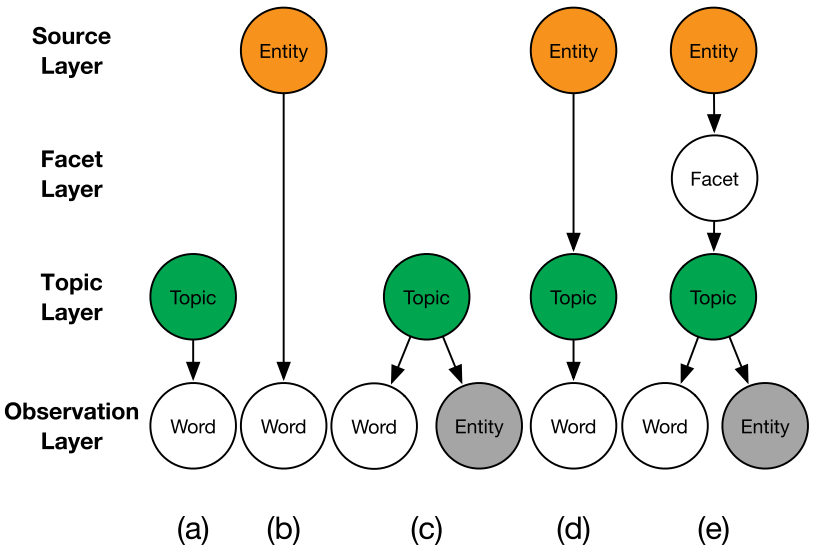

Fig. 2. Different dependencies among words, entities, topics, facets, sources. The simplified plate representations of the models are: (a) LDA; (b) Author Model (AM); (c) Link-LDA; (d) Author Topic Model (ATM); (e) EFTM.

2. Related work

We discuss three lines of related work: document representation, topic modeling, and entity facet mining.

2.1. Document representation

Research on document representation dates back (at least) to the early days of the vector space model, which represents documents as vectors of index terms. Index terms are weighted to capture the importance of a term in describing the content of a particular document (e.g., term frequency) and the discriminative power of a term (e.g., inverse document frequency). Many term weighting methods have been proposed to achieve better document representations, including the widely used TF-IDF method. Bouadjenek, Hacid, Bouzeghoub, and Vakali (2016) propose to integrate social information of users in the index structure of an IR system. The index model provides a Personalized Social Document Representation of each document per user based on his/her activities in a social tagging system. Similar with our work, they start from the intuition that each user has his/her own understanding and point of view of a given document. However, our work differs from their work in that our entity-centric document representations are based on latent entity facets learned from document collections, while theirs are based on associated social annotations of users (entities).

In contrast to representing documents using weighted terms, dense vector representations became popular since the prevalence of LDA (Blei et al., 2003). In LDA, a document is represented by a mixture of latent topics, which are further represented by multinomial distributions of words and shared across all documents. By using topic distributions to represent documents, the topical difference between documents is captured. Extensions of LDA usually model different document generative processes, while using topic distributions to represent documents. Traditionally, a single topic distribution is learned for each document and applied whenever document representations are used. Our work aligns with this line of research and differs from existing work in that we propose to learn multiple entity-centric representations for each document, where each representation corresponds to an entity and is a mixture of facets of the entity.

Today, latent document representations such as doc2vec (Le & Mikolov, 2014) are often inferred using neural networks. Doc2vec is an extension of word2vec (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013) that learns document-level embedding to predict the next word given contexts sampled from a document. Van Gysel, de Rijke, and Kanoulas (2018) propose a neural vector space model (NVSM), which learn document representations directly by gradient descent from sampled n-gram/document pairs extracted from a given corpus. Compared to our proposed representation, doc2vec and NVSM consider neither entities in documents nor entities as sources of information. Latest works on contextualized word representations, such as BERT (Devlin, Chang, Lee, & Toutanova, 2018) and ELMo (Peters et al., 2018), are becoming increasingly popular. The major difference between our work and theirs is that we consider document representations with respect to entities, while they infer word representations with respect to contexts around words.

Recent work has extended the embedding paradigm to entities (Van Gysel, de Rijke, & Kanoulas, 2016a; Van Gysel, de Rijke, & Worring, 2016b); however, their focus is restricted to the semantics of entities, whereas we focus on the relation between entities, entity facets, topics and documents. Xiong, Callan, and Liu (2017) propose a word entity duet representation for ad-hoc retrieval, which use entity based representation of documents. Dai, Tang, Wu, and Zhuang (2018) propose to learn entity mention aware document representation, which learns representations of documents from semantics between not only document-word pairs, but also document-entity pairs and entity-entity pairs. Raviv, Kurland, and Carmel (2016) devise an entity based language model which takes into account both the uncertainty inherent in the entity-markup process and the balance between entity-based and term-based information. Their work differs from ours in that they are using entities in documents as a bag-of-entities representation, while our work considers learning multiple document representations, each of which corresponds to an entity semantically relevant to the document.

2.2. Topic models

LDA (Blei et al., 2003) models a document as a mixture of topics, and automatically generates summaries of topics in terms of a multinomial distribution over words. Many extensions based on it have been proposed to learn topics by assuming alternative document generative process, thus leading to alternative semantics of learned document representations. E.g., Link-LDA (Erosheva et al., 2004) (Fig. 2(c)) extends LDA by modeling words and references (viewed as entities) of an article separately.

In recent years, people focus on extending topic models to address specific tasks, such as short text modeling (Bicalho, Pita, Pedrosa, Lacerda, & Pappa, 2017; Li et al., 2018; Zhang, Mao, & Zeng, 2016), user clustering (Qiu & Shen, 2017), dataless text classification (Li, Xing, Sun, & Ma, 2016), and opinion mining (Ma, Zhang, Liu, Li, & Yuan, 2016). In contrast to designing topic models for specific tasks, our work aims at mining entity facets so as to learn entity-centric document representations, which could be used in downstream applications. Therefore, our work is particularly related to topic models which either considers entities in documents or external labels associated to documents.

The Author Model (AM) (McCallum, 1999) (Fig. 2(b)) and Author Topic Model (ATM) (Rosen-Zvi, Griffiths, Steyvers, & Smyth, 2004) (Fig. 2(d)) have been proposed for multi-labeled documents, where each label (author) is viewed as an entity. In AM, words are generated by first selecting an author and then sampling from an author-specific multinomial distribution over words. ATM extends AM by introducing topics between authors associated with documents and words in documents. In particular, ATM chooses a latent topic from an entity-specific multinomial distribution over topics; a word is then drawn from a topic-specific multinomial

distribution. Similar to AM and ATM, we consider labels of documents as source entities. However, ATM focus on learning representation for entities (authors) and does not learn document representations, while our model aims at learning document representations. There are other entity topic models, such as an entity topic model for entity linking (Han & Sun, 2012), and a hierarchical entity topic model designed for streaming data (Hu, Li, Zhang, & Shao, 2015). Newman, Chemudugunta, and Smyth (2006) propose CorrLDA2 to learn the relationship between topics discussed in news articles and entities mentioned in articles.Chang, Boyd-Graber, and Blei (2009) propose a topic model, which analyze free text to extract descriptions of relationships between entities. These models differ with our work in that they aim at resolving particular tasks, while we focus on modeling entities to learn better document representations.

The Entity Topic Model (ETM) (Kim, Sun, Hockenmaier, & Han, 2012) represents entities in the same way as latent topics. For each document, a topic distribution is drawn from a Dirichlet prior and a joint multinomial distribution over words Φ is obtained by linearly combining entities and topics of a document. To generate a word, a topic is sampled from Φ and a word is sampled from the topic word distribution. Though ETM seems to be a valid baseline for our work, it is not applicable because of scalability issues. In particular, given N words, E entities, F facets and K topics, the number of parameters of ETM is $K \times N + E \times N + E \times K \times N$, which represents latent topics, entity topics and entity-topic pairs, while that of our model is $K \times N + E \times F \times K$, which represents latent topics and entity facets. The number of parameters of ETM increases fast with increasing number of entities because of the component $E \times K \times N$. For example, given 10 entities, 100 topics and 10,000 words, the number of parameters becomes 10 million. In comparison, $E \times F \times K$ is much smaller than $E \times K \times N$ due to the facet that $F \ll N$.

A different line of extensions to LDA focuses on leveraging supervised labels. The supervised topic model in (Mcauliffe & Blei, 2008) is designed for single-labeled documents, while our model is mainly designed for multi-labeled documents. Though also designed for multi-labeled documents, the hierarchically supervised topic model proposed in Perotte, Wood, Elhadad, and Bartlett (2011) considers the scenario where labels of documents form a hierarchy, while our models do not consider label hierarchy. One work considering both supervised and flat labels is Labeled LDA (Ramage, Hall, Nallapati, & Manning, 2009); it focuses more on credit attribution within tagged documents or visual analysis, instead of learning better document representation. The constraint that one label corresponds to one topic helps in their task but limits the representation ability of Labeled LDA in that a label is a high granularity semantic unit that might have various facets. Compared to Labeled LDA, we consider the labels of a document as source entities, which themselves consist of smaller semantic units, i.e., entity facets. The advantage of our model is that different facets related to labels are captured by entity facets, instead of being mixed under the same topic. Other topic models that are relevant to our work include DFLDA (Li, Ouyang, & Zhou, 2015b) and CPTM (Li, Ouyang, & Zhou, 2015a). Since these models are targeted on the task of multi-label classification and make use of global prior information, such as label frequency, we do not consider them as baselines.

Another family of topic models that look similar to our model are topic models with a hierarchy of topics. To generate a document using the hierarchical topic model (Griffiths & Tenenbaum, 2004), a path with L nodes from the root node of a tree to a leaf is selected and a vector of topic proportions θ is drawn from an L-dimensional Dirichlet distribution. Then, words in the document are drawn from a mixture of the topics along the path with mixing proportions θ. Since different nodes in the topic hierarchy represent different topics, the semantics of the topic proportions representation of different documents are different. While somewhat similar to a two-layer hierarchy of topics, our model is different in that the second layer of topics in our model are entity facets (entity specific topics), which makes it specific to entities compared to general topics defined in a two layer hierarchical topic model.

2.3. Entity facet mining

With the growing importance of semantic search and knowledge graphs in recent years, mining and leveraging entity information has received considerable attention (Balog, 2018; Bast, Buchhold, & Haussmann, 2016). Among various categories of information of entities, such as facts and entity relations, entity facets are also considered useful for entity related tasks; Eg. Reinanda, Meij, and de Rijke (2016) use entity aspect similarity as a feature to help filtering documents for long-tail entities.

Significant work in entity facet mining has been conducted in relation to product facets and online reviews. Applications includes product related QA (Yu & Lam, 2018), online review mining (Alam, Ryu, & Lee, 2016; Dragoni, Federici, & Rexha, 2018; Xiao, Ji, Li, Zhuang, & Shi, 2018), review summarization (Liu, Fang, Choulos, Park, & Hu, 2017). Titov and McDonald (2008) propose to first extract ratable facets of objects from online user reviews and then cluster them into coherent topics. Yu, Zha, Wang, and Chua (2011) automatically identify important product (entity) aspects from online consumer reviews. Sikchi, Goyal, and Datta (2016) propose an aspect based product comparator to help consumers in purchasing decision making. Yu and Lam (2018) propose to learn aspects of product categories to predict answers for product-related questions. Li, Wang, Gao, and Jiang (2011) develop an event-aspect topic model to cluster sentences into aspects for events. The related works above focus on particular category of entities, while our work is targeted on mining facets for general purpose entities related to textual documents.

In addition to focusing on particular categories of entities, there are also works on mining facets of general purpose entities from various sources, such as Wikipedia (Fetahu et al., 2015; Nanni, Ponzetto, & Dietz, 2018), query logs (Reinanda, Meij, & de Rijke, 2015) and microblog posts (Spina et al., 2012). Spina et al. (2012) propose to identify entity aspects from social web streams (such as tweets) in the field of online reputation management. Given all queries containing an entity, Reinanda et al. (2015) propose to obtain *query contexts* by removing mentions of the entity in queries. Then all query contexts are clustered and entity facets are identified as clusters which includes similar query contexts. Our work differs from existing works in that we identify entity facets from document collections where documents are associated to entities. The major advantage is that information related to entities can be widely and mostly found in textual collections.

On the other hand, our work differs from existing work in terms of facet representation. In existing works, entity facets are represented by a bag of text segments (Reinanda et al., 2015; Spina et al., 2012), textual description (Nanni et al., 2018), or sentence patterns (Li, Wang, & Jiang, 2013). Spina et al. (2012) consider terms as aspects and try to rank list of aspects that are being discussed with respect to a given company. Li et al. (2013) propose a model to perform clustering, and use the clustered sentences and words as aspects. Nanni et al. (2018) directly use sections in Wikipedia pages as aspects of corresponding entities. In comparison, our work represents entity facets as a mixture of latent topics and use it as the basis of entity-centric representations. The facet representations are useful in existing work with regard to their end goal. However, our choice is advantageous in our case for the following reasons. First, we are not targeted on particular categories of entities, which makes it impossible to simply find sentence patterns by clustering. Second, our end goal is to learn entity-centric representations, which are based on entity facets. By representing entity facets as a mixture of topics, we can jointly learn entity-centric representations and entity facets using our proposed topic model.

Overall, we extend the state-of-the-art in three ways: a new task (learning entity-centric document representations), a new topic model to address this task (the entity facet topic model), and a new way of capturing and reasoning about entity-related facet information.

3. Research objectives

The key objective of our research is to model documents with respect to the specific facets of the entities that are discussed in each document. In particular, our work derives from the hypothesis that entities are not monolithic concepts, but instead have different facets, and documents associated with certain entities discuss these entities from a specific facet perspective. Our goal is to automatically identify entity facets from documents and derive multi-faceted entity-centric representations of the documents in a collection.

We set forth the following research objectives: *RO1 Modeling entity facets, as a mixture of latent topics, and learning them from documents associated to these entities.*

Our first research objective aims to set up a theoretical definition of entity facets based on latent topics. By defining entity facets as a mixture of latent topics, we connect the specificity of entities (entity facets) to the generality of documents (topics in documents).

RO2 Learning multiple entity-centric document representations based on entity facets.

The focus of our second research objective is to model the generative process of documents as a joint effort of particular facets of entities. In this way, we attempt to learn both facets and representations of documents which are based on these facets.

RO3 Confirming that considering entity facets has practical implications in downstream applications.

The focus of the third objective is to understand whether considering entities as multi-faceted concepts makes a difference when it comes to downstream applications, i.e. whether denoising entity and document representation by focusing on specific facets discussed in a document can help in applications such as text classification.

RO4 Identifying by the means of predictive modeling what should one consider as an effective number of facets an average entity has, and how many topics, in traditional terms, are good enough to effectively define these facets.

The focus of the last objective is to gain a better understanding of how many facets an average entity has within a collection of documents. Clearly different entities may have different number of facets, but in this work, we focus on what is the effective number of them on average. Further, one could explore different ways to validate the number of facets. In this work we focus on the predictive power of the multi-faceted entity-centric document representation to fulfil this objective.

Table 1
Notation.

Symbol	Description
D	document collection
V_W	word vocabulary
V_E	document entity set
S	source entity set
$\mathbf{w}_d$	bag of words in document d
$\mathbf{e}_d$	bag of entities in document d
$\mathbf{S}_d$	set of source entities associated with document d
s_d	s-th source entity in d
F	number of entity facets
K	number of topics
f_s	facet f of source entity s
z_s	topic selected from f_s
ϕ_k	word distribution of topic k
φ_k	document entity distribution of topic k
Φ^k	topic token distribution of topic k
η_s	weight of ϕ_s when doing linear combination
$\rho_{f,s}$	facet topic distribution of f_s
θ_s	multinomial distribution of facets of s
$w_{d,\ i}$	the i-th word in document d
$e_{d,j}$	the j-th entity in document d

To address the above objectives, we propose to learn entity facets and entity-centric document representations, where each representation corresponds to an entity and is a mixture of entity facets of the entity.

4. Problem formulation

The task of *entity-centric representation learning* is formulated as follows. Given a collection of documents D, in which each document d consists of a bag of words $\mathbf{w}_d$ and a bag of entities $\mathbf{e}_d$, and is associated with a set of source entities $\mathbf{S}_d$, our goal is to learn entity-centric document representations for all documents in D. The elements in $\mathbf{w}_d$, $\mathbf{e}_d$, and $\mathbf{S}_d$ belong to a word vocabulary V_W, an entity vocabulary V_E, and a source entity set S, respectively. The association between source entities and documents is assumed to be predefined. Table 1 lists the main notation we use.

Here, we define entities as unique identifiers, such as tags of pictures and entities in documents as represented by identifiers in a collaborative knowledge base (e.g., machine IDs in Freebase). Source entities are entities that satisfy the following two conditions: (1) Source entities are topically multi-faceted; (2) each source entity is associated with a group of documents. For example, in Example 1 in Section 1, entities can be viewed as source entities because entities usually have multiple facets (multi-faceted) and each entity is related to many documents that centred around them. Note that source entities are usually different from document entities. For example, tags in news articles, or authors of papers are considered as source entities, while mentions of people or location entities in news articles are document entities. However, the sets of source entities and document entities could also overlap or be identical. For example, if someone wants to learn facet information of Freebase entities in an entity-annotated document collection, they can define the source entities to be the salient entities of a document.

To make matters concrete, Fig. 1 provides an example of the entity-centric representation learning task. Given three documents, d_0, d_1 and d_2, which are associated with source entities {VP}, {BO, VP} and {BO}, respectively, our goal is to learn a document representation of d_0 for VP, of d_1 for VP and BO, and of d_2 for BO, as shown in the figure.

5. Method

In this section, we introduce our method for learning entity-centric document representations. We first provide an overview of how we define entity-centric document representation. Then we propose a novel topic model to model the process of generating documents, which is followed by a Gibbs sampling-based learning algorithm.

5.1. Overview

To define entity-centric document representation, we first introduce the concept of an entity facet. An *entity facet* is a latent aspect of a specific entity. Each facet is represented by a mixed proportion of latent topics. Unlike LDA's topics that are defined as probability distributions over words, each topic in our model is defined as two separate probability distributions, one over words and one over document entities, respectively, which helps to account for the different observed variables.

An entity-centric document representation is defined as a mixed proportion of entity facets, called (entity) facet distributions.

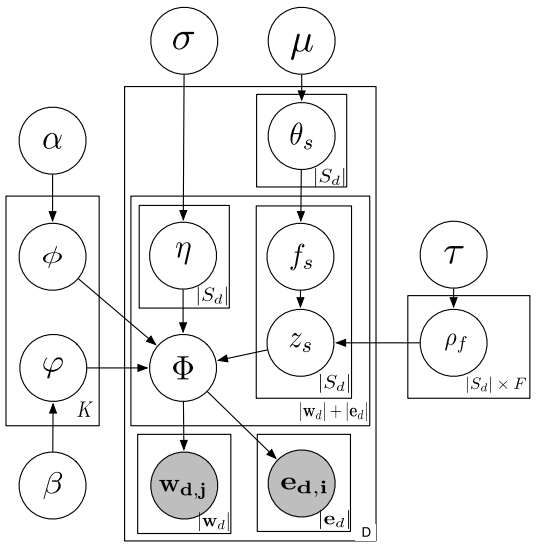

Fig. 3. A graphical representation of the entity facet topic model (EFTM). A concise overview of parameters is as follows: τ, α, β and μ are Dirichlet priors used to generate corresponding multinomial distributions; η is a Beta prior used to generate Binomial distribution; ρ_f is the facet topic distribution of the f-th facet of an entity; ϕ and ψ are topic word distribution and topic (document) entity distribution respectively.

Given a document d associated with a set of source entities $\mathbf{S}_d$, the entity-centric representation of d is a set of facet distributions $\{\theta_s|s \in \mathbf{S}_d\}$. The goal of our model is to learn $\{\theta_s|s \in \mathbf{S}_d\}$ for all d in the document collection D. As part of the model, we also learn: (1) facet topic distributions ρ_f, i.e., a multinomial distribution over topics; (2) topic word distributions ϕ, and (3) topic (document) entity representations φ, i.e., multinomial distributions over words and document entities.

5.2. Entity facet topic model

A graphical representation of the entity facet topic model (EFTM) is shown in Fig. 3; the generative process underlying EFTM is given in Algorithm 1, while detailed explanations are given below.

5.2.1. Generative process

During model initialization, several multinomial distributions are drawn from Dirichlet priors. For each topic, a topic word distribution ϕ and topic entity distribution φ are generated using Dirichlet priors α and β. For facet f of source entity s, the corresponding facet topic distribution $\rho_{f,s}$ is drawn from a Dirichlet prior τ.

In the generative process, given a document d associated with a set of source entities $\mathbf{S}_d$, a set of multinomial distributions $\{\theta_s|s \in \mathbf{S}_d\}$ is generated using a Dirichlet prior μ, where θ_s is a distribution over entity facets of source entity s. To generate a token (word or entity), we iterate over each source entity s in $\mathbf{S}_d$, and draw a facet f_s from θ_s, a topic z_s from $\rho_{f,s}$, and a weight η_s from $B(\sigma)$. Then, for each topic z_s, its topic word distribution and topic entity distribution are first weighted using η_s and then concatenated to obtain a new multinomial distribution Φ_z^s, which is referred to as the *topic token distribution*. In this way, words and entities under the same topic are correlated. Finally, the final topic token distribution Φ is obtained as an equally weighted combination of a set of topic token distributions $\{\Phi_z^s|s \in S_d\}$, which is then used to generate a token, i.e., either a word or entity.

5.2.2. Joint facets

In existing topic models, a document is usually represented by one topic distribution, which is assumed to be used to generate the document. To generate a word, a topic is selected from the topic distribution of the document, and a word is sampled from the topic word distribution of the selected topic. For our model, in order to learn multiple representations for a document, we assume that all source entities contribute to the generation of words and entities in documents. In particular, to generate a word or an entity, a facet is sampled from the facet distribution of each source entity, and a topic is sampled from the facet topic distribution of selected facets. Then, all selected topics are merged by a weighted combination of the corresponding topic word distribution and topic entity distribution, and the resulting topic token distribution is used to sample a word or token. The intuition behind this is that the facet distribution of all source entities of a document should contribute to the generation of words and entities in the document. For example, given a document with three source entities: *USA, Barack Obama*, and *Mitt Romney*, the representation of this document

for each topic z **do**
  Draw $\phi \sim Dir(\alpha)$
  Draw $\varphi \sim Dir(\beta)$
end for
for each source entity $s \in S$ **do**
  Draw $\theta_s \sim Dir(\mu)$
  for each facet of source entity f_s **do**
    Draw $\rho_{f,s} \sim Dir(\tau)$
  end for
end for
for each document d **do**
  for each token t_x **do**
    for each source entity $s \in S_d$ **do**
      Draw $f_s \sim \theta_s$
      Draw $z_s \sim \rho_{f,s}$
      Draw $\eta_s \sim B(\sigma)$
      $\Phi_t^s = \eta_s \phi_{z_s} \oplus (1 - \eta_s)\varphi_{z_s}$
    end for
    $\Phi = \frac{1}{|S_d|} \sum_{s \in S_d} \Phi_t^s$
    Draw $t_x \sim \Phi$, $t_x \in V_W \cup V_E$
  end for
end for

Algorithm 1. Generativeprocess of the entity facet topic model.

could be dependent on a facet that is semantically closest to the "presidential campaign."

5.2.3. Parameters

The number of parameters to be estimated is $K \times (|V_W| + |V_E|) + |S| \times F \times K$, where K is the number of topics, $|S|$ is the size of the source entity set, and F is the number of facets. We use symmetric Dirichlet priors to generate multinomial distributions.

5.2.4. Model advantage

The advantage of our model over existing document representation methods lies in two facets. First, the multi-faceted nature of source entities is incorporated into our model, which captures the rich semantics captured by different facets of source entities. Second, entity specific facets and general aspects (topics) are semantically connected, thus bridging the gap between entities and topics.

6. Inference for EFTM

In this section, we present our Gibbs sampling algorithm for EFTM. Two key latent variables used to generate a word w_i are estimated, i.e., z_s^i and f_s^i. The former is the topic sampled from a given facet f_s, while the latter is the facet selected from a given source entity s. The parameter estimation process is described as follows.

The conditional posterior distribution for z_s^i, the topic from the f-th facet of source entity s, to generate word w_i is

$$P(z_s^i = t|\mathbf{z}_{-s}^i, \mathbf{z}_{(\cdot)}^{-i}, \mathbf{f}, \mathbf{w}, \mathbf{e}) \propto P(w_i|z_s^i = t, \mathbf{z}_{-s}^i, \mathbf{w}_{-i}, \mathbf{e}) \cdot P(z_s^i = t|\mathbf{z}_{-s}^i, \mathbf{z}_{(\cdot)}^{-i}, \mathbf{f}), \tag{1}$$

where $\mathbf{z}_{-s}^i$ is the assignment of all z related to word w_i except for z_s^i, $\mathbf{z}_{(\cdot)}^{-i}$ is the assignment of all z for all words and entities except for w_i, $\mathbf{f}$ is the assignment of all facets for all words and entities, $\mathbf{w}_{-i}$ are all words except word w_i, $\mathbf{e}$ are all entities.

For the first item on the right hand side of Eq. 1, we have

$$P(w_i|z_s^i = t, \mathbf{z}_{-s}^i, \mathbf{w}_{-i}, \mathbf{e}) = \int P(w_i|z_s^i = t, \mathbf{z}_{-s}^i, \Phi) P(\Phi|\mathbf{z}_{(\cdot)}^{-i}, \mathbf{w}_{-i}, \mathbf{e}) d\Phi. \tag{2}$$

Here, Φ is the joint word and entity distribution over all facets of source entities. In particular, given a $|V_W|$-dimensional word distribution and a $|V_E|$-dimensional entity distribution under topic t, a $|V_W| + |V_E|$-dimensional joint multinomial distribution over words and entities Φ_t can be obtained as a weighted concatenation. For the topic t sampled from a facet of source entity s, the topic token distribution Φ_t^s is obtained as a weighted concatenation of ϕ_t and φ_t with weights η_s and $(1 - \eta_s)$, respectively. Since different source entities are assumed to contribute equally, Φ is obtained as follows:

$$\Phi = \frac{1}{|S_d|} \sum_{s \in [1, |S_d|]} \Phi_t^s = \frac{1}{|S_d|} \sum_{s \in [1, |S_d|]} (\eta_s \phi_t \oplus (1 - \eta_s)\varphi_t), \tag{3}$$

where $\oplus$ means concatenation of two distributions.

Based on how we obtain Φ, we rewrite Eq. 2 as follows:

$$P(w_i|z_s^i = t, \mathbf{z}_{-s}^i, \mathbf{w}_{-i}, \mathbf{e}) = \frac{1}{|S_d|} \left(P(w_i|z_s^i = t, \mathbf{w}_{-i}, \mathbf{e}) + \sum_{s' \in S_d \setminus \{s\}} P(w_i|z_{s'}^i) \right), \tag{4}$$

where $P(w_i|z_{s'}^i)$ is $n_{i,s'}^{w_i}$, the number of instances of w_i under topic $z_{s'}^i$. For the first item in Eq. 4, we have:

$$P(w_i|z_s^i = t, \mathbf{w}_{-i}, \mathbf{e}) = \int P(w_i|z_s^i = t, \Phi_t) P(\Phi_t|\mathbf{w}_{-i}, \mathbf{e}) d\Phi_t. \tag{5}$$

For the second item on the right-hand side of Eq. 5, we have:

$$P(\Phi_t|\mathbf{w}_{-i}, \mathbf{e}) \propto P(\mathbf{w}_{-i}, \mathbf{e}|\Phi_t) P(\Phi_t). \tag{6}$$

Since $P(\phi_t)$ and $P(\varphi_t)$ are Dirichlet priors $Dir(\alpha)$ and $Dir(\beta)$, and $P(\eta_s)$ is $Beta(\sigma)$, the prior distribution $P(\Phi_t)$ is:

$$\frac{\sigma_1}{\sigma_1 + \sigma_2}\alpha + \frac{\sigma_2}{\sigma_1 + \sigma_2}\beta,$$

where $\sigma = [\sigma_1, \sigma_2]$ is a Beta prior. Since Φ_t is conjugate to the likelihood function (the first item in Eq. 5), the posterior distribution in Eq. 5 is obtained by putting the multinomial likelihood into the Dirichlet prior:

$$Dir\left(\frac{\sigma_1}{\sigma_1 + \sigma_2}\alpha + \frac{\sigma_2}{\sigma_1 + \sigma_2}\beta + n_{-i,t}^{w_i} \right),$$

where $n_{-i,t}^{w_i}$ is the number of words and entities that is assigned to Φ_t. Combining the previous equations, we have:

$$P(w_i|\ z_s^i = t,\ \mathbf{z}_{-s}^i,\ \mathbf{f},\ \mathbf{w}_{-i},\ \mathbf{e}) = \frac{1}{S_{|d|}}\left(\frac{\sum_{s' \in S_d \setminus \{s\}} n_{i,s'}^{w_i}}{\sum_{s' \in S_d \setminus \{s\}} n_{i,s'}^{(\cdot)}} + \frac{(\sigma_1+\sigma_2)\cdot n_{-i,t}^{w_i} + \sigma_1\alpha + \sigma_2\beta}{(\sigma_1+\sigma_2)\cdot n_{-i,t}^{(\cdot)} + (\sigma_1\alpha+\sigma_2\beta)\cdot(|V_E|+|V_W|)}\right). \tag{7}$$

This addresses the first term on the right-hand side of Eq. 1.

For the second term on the right-hand side of Eq. 1, we have:

$$P(\ z_s^i = t|\mathbf{z}_{-s}^i,\ \mathbf{z}_{(\cdot)}^{-i},\mathbf{f}) = \int P(z_s^i = t|f_s^i,\ \rho_{f,s})P(\rho_{f,s}|\mathbf{z}_s^{-i},\ \mathbf{f}_s^{-i})d\rho_{f,s} \tag{8}$$

The second item on the right-hand of Eq. 8 is a posterior as follows:

$$P(\rho_{f,s}|\mathbf{z}_s^{-i},\ \mathbf{f}_s^{-i}) \propto P(\mathbf{z}_s^{-i}|\rho_{f,s},\ \mathbf{f}_s^{-i})P(\rho_{f,s}). \tag{9}$$

Here, $P(\rho_{f,s})$ is a Dirichlet prior $Dir(\tau)$ and $P(\mathbf{z}_s^{-i}|\rho_{f,s},\ \mathbf{f}_s^{-i})$ is the number of all words and entities that are assigned to topic t because of source entity s except w_i, denoted as $n_{s,t}^{-i}$. Then, Eq. 8 can be written as:

$$P(z_s^i = t|\mathbf{z}_{-s}^i,\ \mathbf{z}_{(\cdot)}^{-i},\ \mathbf{f}) = \frac{n_{s,t}^{-i} + \tau}{n_{s,\cdot}^{-i} + K\tau}. \tag{10}$$

The conditional posterior distribution for f_s^i, the facet chosen to first generate a topic and then generate w_i, is:

$$P(f_s^i = f|\mathbf{f}_{-s}^i,\ \mathbf{f}_{(\cdot)}^{-i},\ \mathbf{z}) \propto P(z_s^i|f_s^i = f,\ \mathbf{z}_s^{-i})P(f_s^i = f|\mathbf{f}_s^{-i}). \tag{11}$$

For the first item on the right-hand side of Eq. 11, we have:

$$P(z_s^i|f_s^i = f,\ \mathbf{z}_s^{-i}) = \int P(z_s^i|f_s^i = f,\ \rho_{f,s})P(\rho_{f,s}|\mathbf{z}_s^{-i})d\rho_{f,s}. \tag{12}$$

The second item on the right hand side of Eq. 12 is:

$$P(\rho_{f,s}|\mathbf{z}_s^{-i}) \propto P(\mathbf{z}_s^{-i}|\rho_{f,s})P(\rho_{f,s}). \tag{13}$$

Since $P(z_s^{-i}|\rho_{f,s})$ is a likelihood function and $P(\rho_{f,s})$ is a Dirichlet prior $Dir(\tau)$, we have $P(f_s)$ as $Dir(\tau + n_{f,s}^{-i})$, where $n_{f,s}^{-i}$ is the number of topics assigned to f_s except for the current word.

We combine all equations and obtain:

$$P(z_s^i|f_s^i = f,\ \mathbf{z}_s^{-i}) = \frac{n_{f,s}^i + \tau}{n_{\cdot,s}^i + K\tau}. \tag{14}$$

The second term in Eq. 11 is obtained as follows:

$$P(f_s^i = f|\mathbf{f}_{-s}^{(\cdot)}) = \int P(f_s^i = f|\theta_s)P(\theta_s|\mathbf{f}_s^{-i})d\theta_s. \tag{15}$$

The second term on the right-hand side of Eq. 15 is as follows:

$$P(\theta_s|\mathbf{f}_s^{-i}) \propto P(\mathbf{f}_s^{-i}|\theta_s)P(\theta_s). \tag{16}$$

As a result, we have Eq. 15 as follows:

$$P(f_s^i = f|\mathbf{f}_{-s}^{(\cdot)}) = \frac{n_f^{-i} + \mu}{n_{(\cdot)}^{-i} + F\mu}. \tag{17}$$

Finally, by combining Eqs. 1, 7, 10 and 11,14,17, we obtain the desired estimates of posterior distributions of z_s^i and f_s^i, respectively:

$$P(z_s^i = t|\mathbf{z}_{-s}^i,\ \mathbf{z}_{(\cdot)}^{-i},\ \mathbf{f},\ \mathbf{w},\ \mathbf{e}) \propto \frac{1}{S_{|d|}}\left(\frac{\sum_{s' \in S_d \setminus \{s\}} n_{i,s'}^{w_i}}{\sum_{s' \in S_d \setminus \{s\}} n_{i,s'}^{(\cdot)}} + \frac{(\sigma_1+\sigma_2)\cdot n_{-i,t}^{w_i} + \sigma_1\alpha + \sigma_2\beta}{(\sigma_1+\sigma_2)\cdot n_{-i,t}^{(\cdot)} + (\sigma_1\alpha+\sigma_2\beta)\cdot(|V_E|+|V_W|)}\right) \cdot \frac{n_{s,t}^{-i} + \tau}{n_{s,\cdot}^{-i} + K\tau}. \tag{18}$$

$$P(f_s^i = f|\mathbf{f}_{-s}^i,\ \mathbf{f}_{(\cdot)}^{-i},\ \mathbf{z}) \propto \frac{n_{f,s}^i + \tau}{n_{\cdot,s}^i + K\tau} \cdot \frac{n_f^{-i} + \mu}{n_{(\cdot)}^{-i} + F\mu}. \tag{19}$$

7. Experimental setup

We set up experiments to address the following research questions:

RQ1 Can entity-centric document representation capture entity facets reflected in documents?
RQ2 Is entity facet topic model better than existing topic models involving entities in terms of generative capability?
RQ3 Is the multi-faceted entity-centric document representation better than state-of-the-art document representations when used for multi-label classification?
RQ4 What is the expected number of entity facets (and topics) in a document collection?

7.1. Datasets

Two datasets are used in our experiments. Both of them are extracted from the New York Times Corpus (Sandhaus, 2008). Since we consider both words and entities in documents, we use the entity annotations of New York Times articles from 2003–2007 provided by Google (Dunietz & Gillick, 2014). Documents from 2003–2006 are used as a training set, while documents in 2007 are used as the test set. Articles in the New York Times Corpus are all associated with multiple labels, called *descriptors*. Since a source entity is a general concept referring to any semantic unit representing a source of information, descriptors are considered as source entities in EFTM.

To examine the performance of EFTM on datasets of different sizes, we extract two datasets according to the following procedure: We first count the frequency of the descriptors, and then select target descriptors based on their frequency. The statistics about descriptors are presented in Fig. 4. We do not consider descriptors with either high or low number of associated documents, so that our extracted datasets are under moderate size. Articles associated to at least two target descriptors are selected to be part of the output dataset. For the first dataset ("dataset1" in Table 2), we select descriptors whose frequency ranges from 500 to 1,000, which leads to a set of 56 descriptors. For the second dataset ("dataset2" in Table 2), we choose descriptors whose frequency is higher than 2,000, with 33 descriptors being selected.

We count the descriptive statistics of our two datasets and present them in Table 2. The definitions of label cardinality and label density (Tsoumakas & Katakis, 2007) are as follows. Let D be a multi-label dataset consisting of D multi-label examples (x_i, Y_i), $i = 1...D$. Label cardinality of D is defined as the average number of labels of the examples in D:

$$LC(D) = \frac{1}{|D|}\sum_{i=1}^{|D|}|Y_i|$$

Label density of D is defined as the average number of labels of the examples in D divided by L:

$$LC(D) = \frac{1}{|D|}\sum_{i=1}^{|D|}|Y_i|/|L|$$

where L is the number of all labels.

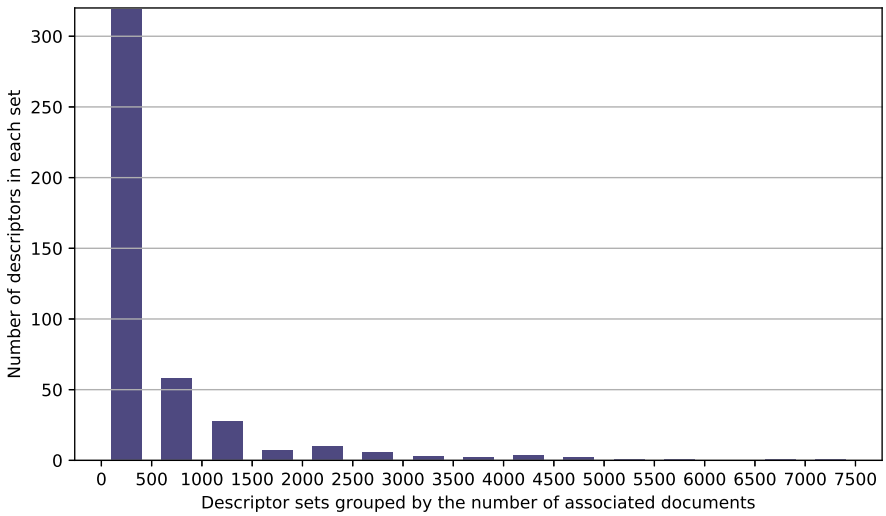

Fig. 4. Statistics of descriptors in NYT Corpus. The x-axis are descriptor sets grouped by the number of associated documents; the y-axis is the number of descriptors in each set.

Table 2
Statistics of our two datasets. Columns 2 and column 3 are the number of documents in the training and test set; column 4 is the number of labels in the dataset; columns 5 and 6 are the label cardinality (LC) and label density (LD).

Dataset	# training	# test	# Labels	LD	LC
1	6377	540	56	0.038	2.139
2	34,976	3146	33	0.077	2.533

7.2. Experimental design

7.2.1. Modeling and recovering entity facets

To answer **RQ1** we conduct a qualitative analysis that seeks to uncover the correlation between the facets of source entities. Our hypothesis is that in a set of documents associated with a pair of source entities, the facets of these entities that are mainly reflected in the documents will be represented by similar topic distributions. If this is true, it would mean that our model is able to identify matching facets of different entities in order to semantically represent a document and it will confirm that indeed entities are not single-faceted. For example, given a set of documents associated with both *Finance* and *Education and Schools*, we expect to identify the funding (education investment) facet of *Education and Schools* and *Finance*. Therefore, given two source entities, first the documents associated to both entities are selected. For each source entity, the corresponding entity-centric representation is computed as the mean of the entity representations of all shared documents. The resulting facet distribution can be considered as a centroid across these documents. Then we examine whether certain facets have a high probability in the facet distribution.

7.2.2. Generative capability of entity facet topic model

To answer **RQ2** we perform a quantitative analysis, using perplexity as the evaluation measure. Perplexity is a standard measure for estimating the performance of a probabilistic model. We evaluate EFTM by estimating the perplexity of unseen held-out documents given training documents. A better model will have a lower perplexity of held-out documents on average. We follow the perplexity definition in Kim et al. (2012), which is defined as follows:

$$\exp\left(-\frac{\log P(D^{test}|D^{train})}{\sum_{d\in D^{test}} N_d}\right).$$

Let Φ denote the set of all parameters in a topic model; then

$$P(D^{test}|D^{train}) = \int P(D^{test}|\Phi)P(\Phi|D^{train})d\Phi.$$

This integral can be approximated by averaging $P(D^{test}|\Phi)$ under samples from $P(\Phi|D^{train})$. Note that EFTM can be seen as a generative process that generates words and entities for a given set of source entities. Thus, $P(D^{test}|\Phi)$ is defined as follows:

$$P(D^{test}|\Phi) = \Pi_{d\in D^{test}} P(w_d, e_d|S_d, \Phi).$$

If our model demonstrates lower perplexity it will confirm that indeed a document is better represented and understood by considering the specific facets of an entity discussed in this document.

7.2.3. Multi-label text classification

To answer **RQ3** we perform an extrinsic evaluation of EFTM, for which we consider a multi-label classification task. We assume that a good document representation model should encode distinctive information about whether a document should be associated to a label or not. Therefore, better document representations should yield better performance in multi-label classification.

A simple and straightforward method for multi-class classification is to train a binary classifier for each label using the representation of a document as features. We adopt this setting and use SVM as the binary classifier in our experiments. We have experimented with a number of different classifiers and all of them support our conclusions. Since, the goal of this work is to compare representations and not classification algorithms, we do not report the performance using other classifiers, such as Random Forests, or Naive Bayes. The features are the elements of the vector representations, and the number of features is the dimensionality of the vectors.

Given a label (i.e., a source entity), EFTM provides an entity centric representation of the document corresponding to that label. However, for the purpose of training we also need an entity centric representation of documents that are not associated to this label, to be used as negative instances. *Pseudo inference* (Rao & Wu, 2010) is performed on the document and the label, by assuming that the label is the only source entity associated to the given document, hence inferring the entity-centric document representation using EFTM.

For each label s, we collect N^s_{train} documents associated to the label as positive instances for training. Then we randomly sample N^s_{train} negative instances from documents not associated to this label for training. During testing, for each label and test instance, we perform pseudo inference on the test instance, and input the obtained representation to the binary classifier corresponding to the label. Then, the binary classifier outputs 1 if the test instance is considered to be associated to the label, and 0 otherwise.

To assess the performance of multi-label classification we use three evaluation measures (Bielza, Li, & Larranaga, 2011;

Tsoumakas & Katakis, 2007): *Multi-label accuracy, macro F1*, and *micro F1*.[2] We conduct a statistical significance test in our experiments via a paired *t*-test. Our experiments consist of two steps, i.e. model training and multiple binary classifications. Since we focus on comparing representations, it is the model training step that is considered as the source of uncertainty. We train each model 5 times, and repeat the classification steps to obtain multiple results. All significance tests are performed at $\alpha = .05$ level.

If our model demonstrates a better performance than baseline methods it will confirm that multi-faceted entity-centric representation of documents has a positive effect not only towards better understanding documents but also in downstream applications.

7.2.4. Number of entity facets and topics in a collection

To answer **RQ4** we perform a quantitative analysis similar to previous section altering the number of topics and the number of entity facets in our model and quantifying the effect in terms of the performance of the model in the multi-label classification task.

7.3. Baselines and model variations

Table 3 lists the document representation methods considered in our experiments. In both our intrinsic and extrinsic evaluation we seek to assess the quality of document representations, and not the ability of a machine learning algorithm to succeed in multi-label classification. Our baselines include the traditional bag of words with TF-IDF weights representation (BoW) (Salton, Yang, & Yu, 1975); the LDA (Blei et al., 2003), LLDA (Erosheva et al., 2004) and Labeled-LDA (Ramage et al., 2009) topic model document representations, that allow us to directly assess the quality of EFTM; and Mean Word Embedding (MWE) (Kraft, Jain, & Rush, 2016) and Doc2vec (Le & Mikolov, 2014) as state-of-the-art dense vector representations, similar with the setup in (Van Gysel et al., 2018). The word embeddings used here are 50d GloVE vectors (Pennington, Socher, & Manning, 2014) pre-trained using Wikipedia and Gigaword 5 [3]. We use a pre-trained BERT model (Devlin et al., 2018) with a hidden layer size of 768, 12 Transformer blocks and 12 self-attention heads. Since the representations learned by baseline methods are not optimised for multi-label classification, we use bert-as-service[4] to get the representation of an input sequence without fine-tuning with respect to multi-label classification to make results comparable to other baseline methods, as well as our work. We truncate the input sequences to 512 tokens to meet the restriction of maximum sequence length required by BERT. To get a feeling of how BERT performs on multi-label classification when fine-tuning, we also fine-tune the same pre-trained BERT model for the task of multi-label classification, with a batch size of 32, max sequence length of 512, learning rate of 5e-5. The maximum number of training epochs is set to 3 according to the parameter setup in Devlin et al. (2018).

Note that even though ETM (Kim et al., 2012) is related to our work, it is not a valid baseline, as explained in Section 2. For baseline models that do not distinguish entities from words, we feed unique identifiers of entities together with words, so that our model does not benefit from being fed more data than baselines.

For EFTM, we consider different source entity and observed variables settings. In terms of source entity, we consider a multi-source (MS) setting, where each document is associated with multiple source entities, and a single-source (SS) setting, where all documents are assumed to be associated with one universal source entity. In terms of observed variables, we consider a word-only (WO) setting, which is a simplified version of EFTM in which we only use words as observed variables, and a words and entities (WE) setting, which is the full EFTM model. We write EFTMWO-MS, EFTMWO-SS, EFTMWE-SS and EFTMWE-MS, respectively, to denote these variants. See Table 3 for a summary.

7.4. Parameter settings

Following standard practice (Kim et al., 2012), we set the hyperparameters of the baseline methods and EFTM to pre-defined values. In LDA, LLDA and EFTM, we set both α and β as 0.1. In LLDA and our model, σ is set to 0.5, which means no prior information is known. Note that our model might perform better if σ is set to a value that corresponds to the frequency of words vs. entities. We use uninformative prior 0.5 and leave the impact of σ as future work. In EFTM, we set τ and μ to 0.1. The number of iterations of Gibbs Sampling is set to 2000 for all models. We set the number of topics to {10, 15, 20, 30, 40, 50, 80, 100}, and the number of facets to {5, 10}, so as to be able to compare the effectiveness under different parameter settings.

In this work we make the assumption that all entities have an equal number of facets. Clearly this can not the case with the number of facets most likely changing across entities. Nonparametric Bayesian models could capture this and we leave it for future work. Instead, we fixed the number of facets to 5 (or 10). Remember that the number of facets is a modeling choice and should be decided by empirical evidence. Our choice is derived from the fact that the median number of sections of English Wikipedia articles (which can be viewed as entity facets (Nanni et al., 2018)) is 4 for the entire collection and 7 for a high quality sub-collection (Piccardi, Catasta, Zia, & West, 2018). Also note that the number of facets does not have to be accurate for all entities, as

[2] Bielza et al. (2011) also define Mean Accuracy as an effectiveness measure. However, our data is rather skewed with respect to each label, i.e., for each label there is a small fraction of documents that are associated with it. This allows a naive classifier that predicts each instance not being associated with a label to achieve high performance when measured by Mean Accuracy. For instance, such a classifier, when applied to "Dataset1" achieves a Mean Accuracy of 0.9629. Multi-label Accuracy, macro and micro F1 avoid this bias.

[3] https://catalog.ldc.upenn.edu/LDC2011T07 .

[4] https://github.com/hanxiao/bert-as-service .

Table 3
Methods and baselines used for comparison.

Acronym	Description	Reference
EFTMWO-SS	EFTM with a single and same source entity, and words as the only observed variables.	§4.2
EFTMWO-MS	EFTM with multiple source entities, and words as the only observed variables.	§4.2
EFTMWE-SS	EFTM with a single and same source entity, and considers two kinds of observed variables.	§4.2
EFTMWE-MS	EFTM with multiple source entities, and considers two kinds of observed variables.	§4.2
BoW	Bag of words weighted by TF-IDF; a traditional document representation.	Salton et al. (1975)
LDA	Latent Dirichlet Allocation, which learns a latent topic distribution to represent documents with; a widely used document representation method.	Blei et al. (2003)
LLDA	Mixed membership of topics, which is similar to LDA, except that it considers words and entities separately.	Erosheva et al. (2004)
Labeled LDA	A supervised topic model, which extends LDA to leverage supervised labels of documents.	Ramage et al. (2009)
Doc2vec	Dense vector representation, which is the state-of-the-art neural method for learning document representations.	Le and Mikolov (2014)
BERT	A state-of-the-art vector representation method.	Devlin et al. (2018)
MWE	Mean word embedding, which is a strong state-of-the-art neural method for representing documents.	Chen (2017); Le and Mikolov (2014)
Most-Frequent	A naive baseline which always predict the most frequent label in train set.	-

long as learned facets could be helpful for downstream applications, e.g. classification, as demonstrated in our experiments. A similar scenario with the number of facets for our model is the number of topics for topic models, where the number of topics can be predefined according to empirical evidence (rather than accurate estimation), as long as the learned topics are useful for inferring meaningful topic distributions of documents.

Regarding runtime, we run our model on single core Intel Xeon CPU with 256GB RAM. We run our models under different setups and against two datasets and the training time are presented in Table 4.

8. Results

In this section, we present the outcomes of our experiments and provide answers for our research questions.

8.1. Modeling and recovering entity facets

To answer **RQ1**, we perform a qualitative analysis of the outcomes of the EFTMWE-MS model trained on Dataset 2 with the number of facets set to 5 and the number of topics set to 50. The setup of the qualitative analysis can be found in Section 7.2.1. We choose three pairs of source entities for analysis, i.e., *Finances, Law and Legislation*, and *Education and Schools*, and the results are shown in Fig. 5. As shown in Fig. 5(c), facet 1 of *Education and Schools* has a higher probability, indicating that facet 1 is related to the law facet of *Education and Schools*, while facet 3 of *Law and Legislation* is related to its education facets. Further, we see that in Fig. 5(a) and 5(b), while there is no particular facet with a strong presence in these documents, there is a reverse relation between facet 1 and 2 of *Finances*. When documents are associated with *Education and Schools*, facet 2 of *Finances* has a higher presence than 1, while the opposite is true when documents are associated with *Law and Legislation*. Hence, different facets of source entities are captured by the entity facets proposed in EFTM.

To further investigate whether these facets indeed capture similar information, we present the facet topic distribution of those facets, i.e., facet 4 of *Finances*, facet 1 of *Education and Schools*, facet 3 of *Law and Legislation* in Fig. 6. The topic distributions of all the facets we consider are similar, demonstrating that similar information is captured by entity facets and understood from the perspective of source entities.

Therefore, the document representations we propose are able to discover and model facets of different entities within a document that semantically align with each other and the theme of the document, while at the same time disagree with the facets learnt by documents of different theme.

Table 4
The number of hours used to train our models under different setups.

	EFTMWO-SS	EFTMWO-MS	EFTMWE-SS	EFTMWE-MS
Dataset 1	4.0 h	9.2 h	3.9 h	9.0 h
Dataset 2	13.2 h	42.9 h	12.6 h	38.9 h

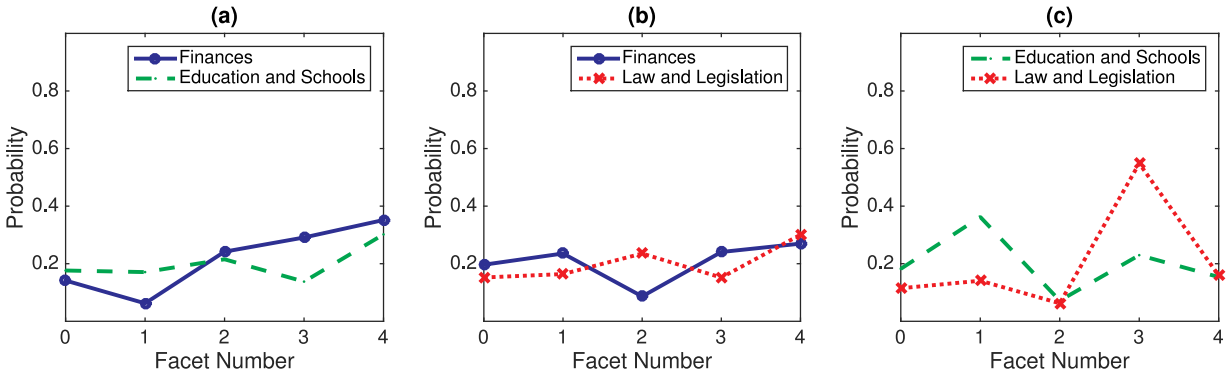

Fig. 5. Analysis on three sets of documents, where each set of documents is associated with the same pair of source entities. Each line is the averaged facet distribution of a specific source entity in the corresponding documents. The x-axis is the facet number, while the y-axis is the value of elements in the facet distributions. The number of facets is 5.

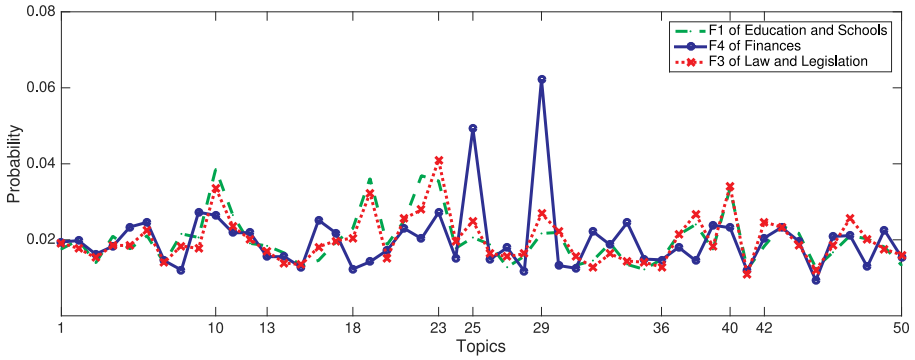

Fig. 6. Facet topic distribution of facet number 1, 4, 3 of *Education and Schools, Finances* and *Law and Legislation*, respectively.

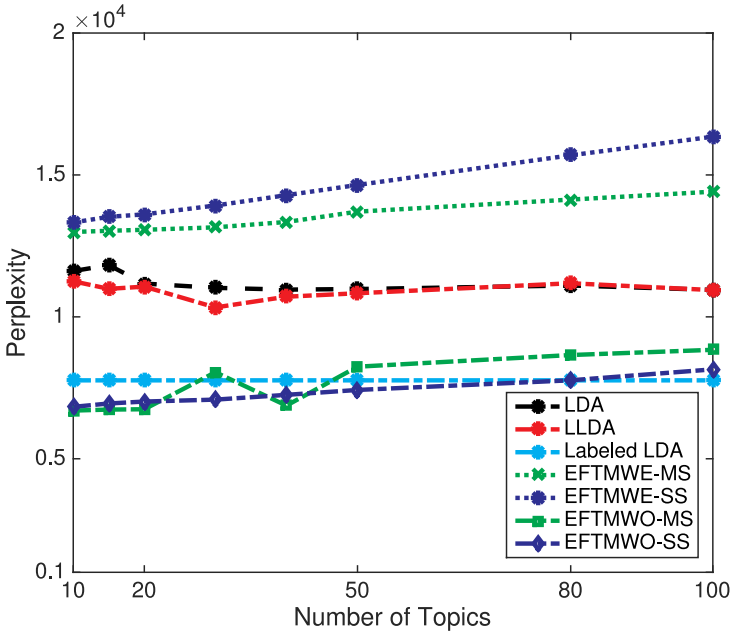

Fig. 7. Perplexity values on Dataset 1 with different numbers of topics.

Finding 1
Entities are associated with documents by means of specific facets of them discussed in these documents, while different documents may focus on different facets of an entity; this confirms our hypothesis that entities should be considered multi-faceted concepts.

8.2. Generative capability of entity facet topic model

To answer **RQ2**, we compare the perplexity values obtained by the different topic models. The perplexity scores over the two datasets are shown in Figs. 7 and 8, respectively. Note that the number of topics of Labeled-LDA is decided by the number of labels, thus its perplexity value is fixed. EFTMWO-MS and EFTMWO-SS perform better than the baseline methods, whereas EFTMWE-MS and EFTMWE-SS perform worse. The difference lies in the representation of topics. In the WE-version of EFTM, topics are represented by two distributions, i.e., a topic-word distribution and topic-(document) entity distribution. The prior of σ is set to 0.5, which means that words and entities have the same chance of being selected to generate a document. However, in our dataset, there are much more words than entities, which leads to a lower probability of generating the documents and higher perplexity values. By introducing entity facets, we achieve better generative capability for unseen documents.

Finding 2
The generative capability of a model that considers entities as multi-faceted concepts is better compared to models that do not.

8.3. Multi-label text classification

To answer **RQ3**, we present the performance of different document representation methods on the multi-label classification task. Results over two datasets are shown in Tables 5 and 6.

As we can see, semantic representation methods, such as LDA, MWE, D2V and BERT, perform better than traditional BoW representations. When comparing semantic representation methods, BERT is better than LDA but not as good as advanced topic models, such as LLDA, indicating that advanced topic modeling has the potential to learn good representations. As mentioned earlier we also fine-tuned BERT on the classification task. We did that for the small Dataset 1, since fine-tuning for the large dataset (Dataset 2) proved to be a rather computationally expensive task (the task was abandoned after 8 days of training). The BERT-fine-tuned model obtained a ml-Acc of 0.0593, Micro F1 of 0.1082, and Macro F1 of 0.0163, better than many unsupervised representation learning methods, but still not better than the unsupervised representation learning of our advanced topic modeling. Note that the

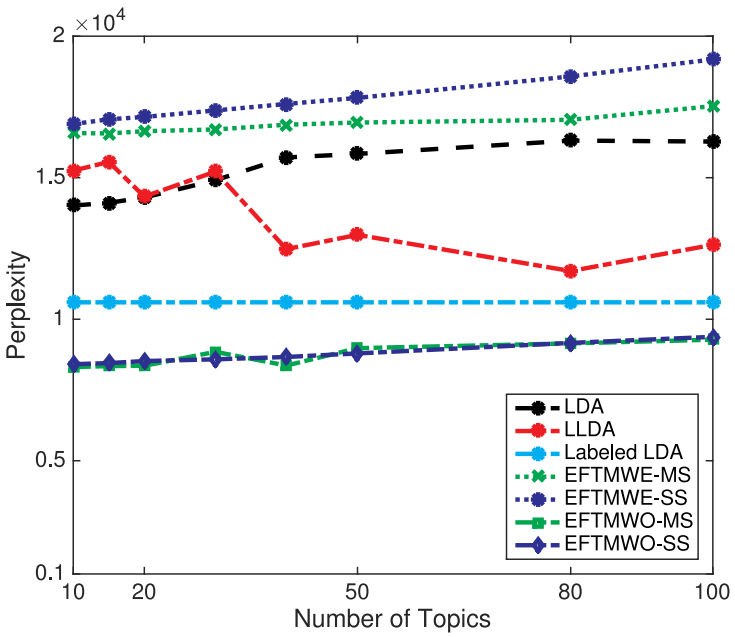

Fig. 8. Perplexity values on Dataset 2 with different numbers of topics.

Table 5
Comparing the performance of document representation methods on the task of multi-label classification on Dataset 1. The number of facets of EFTM is set to 5, and the number of topics/dimensions is set to 50. We test the significance of results of our models compared to baseline methods. Results of our model are significant compared to baseline methods at $\alpha = .05$ level.

	ml-Acc	Micro F1	Macro F1
Most-Frequent	0.0398	0.0571	0.0029
BoW	0.0135	0.0267	0.0272
D2V	0.0047	0.0091	0.0090
MWE	0.0047	0.0094	0.0097
BERT	0.0054	0.0106	0.0109
LDA	0.0064	0.0123	0.0128
LLDA	0.0358	0.0690	0.0632
Labeled LDA	0.0000	0.0033	0.0034
EFTMWO-SS	0.0377	0.0717	0.0710
EFTMWO-MS	0.0392	0.0745	0.0778
EFTMWE-SS	0.0390	0.0738	0.0000
EFTMWE-MS	**0.0399**	**0.0758**	**0.0788**

Table 6
Comparing the performance of document representation methods on the task of multi-label classification on Dataset 2. The number of facets of EFTM is set to 5, and the number of topics/dimensions is set to 50. We test the significance of results of our models compared to baseline methods. Results of our model are significant compared to baseline methods at $\alpha = .05$ level.

	ml-Acc	Micro F1	Macro F1
Most-Frequent	0.0809	0.0983	0.0085
BoW	0.0234	0.0454	0.0478
D2V	0.0088	0.0173	0.0176
MWE	0.0082	0.0161	0.0171
BERT	0.0065	0.0125	0.0133
LDA	0.0082	0.0158	0.0170
LLDA	0.0620	0.1158	0.0948
Labeled LDA	0.0035	0.0069	0.0073
EFTMWO-SS	0.0684	0.1235	0.1185
EFTMWO-MS	0.0658	0.1216	0.0000
EFTMWE-SS	**0.0758**	**0.1387**	**0.1375**
EFTMWE-MS	0.0658	0.1207	0.1214

naive baseline *Most-Frequent* is a strong baseline in terms of multi-label accuracy. This is because that if we predict the most frequent label, it is likely to get at least one right label for each instance. Together with the fact that the label cardinality is around 2, there should be quite a few instances that get almost correctly predicted.

On both datasets, EFTMWO-SS outperforms LDA, indicating the effectiveness of introducing entity facets and representing documents using facet distributions, especially given that the number of features of LDA in multi-label classification (i.e., the number of topics, 50) is much more than that of EFTMWO-SS (number of facets, 5). Hence, representing documents in an entity-centric fashion gives an explicit way to facilitate downstream entity related tasks, such as judging whether a document should be associated with the entity (label) here.

We also study whether it is helpful to distinguish different kinds of observed variables, such as words and entities. The performance of LLDA is better than LDA, showing the effectiveness of distinguishing different kinds of observations in relatively simple topic models. As to variants of EFTM, the performance of EFTMWE-SS/EFTMWE-MS are consistently better than that of EFTMWO-SS/EFTMWO-MS, which confirmed the superiority of considering both observed variables. Note that sometimes macro F1 could be zero because of skewed performance across different labels. Specifically, the performance of many labels are close to zero except few other labels.

To study the impact of multiple sources, we consider two pairs for comparison, i.e., EFTMWE-MS vs. EFTMWE-SS and EFTMWO-MS vs. EFTMWO-SS. On Dataset 1, EFTM with multi-sources (MS) is consistently better than the single-source (SS) version of EFTM, while EFTMWE-SS is better than EFTMWE-MS on Dataset 2. This appears to be related to the size of the set of labels. The number of labels in Dataset 2 is smaller than that of Dataset 1, which lessens the impact of modeling multiple sources.

In sum, EFTMWE-MS and EFTMWE-SS are the preferred choices for the multi-label classification task, with a slight preference for EFTMWE-MS in case where a dataset has many labels (source entities) and for EFTMWE-SS in case the dataset has fewer labels.

Finding 3
By fixing supervised learning algorithm and using different document representations for multi-label classification, we demonstrate that the proposed multi-faceted entity-centric representation outperforms state-of-the-art representations. The smaller the dataset the more important it is to use a multi-source representation.

8.4. Number of entity facets and topics in a collection

To answer **RQ4**, we explore different parameter settings for EFTM. In particular, we consider two cases: (1) Varying the number of topics (30, 50, 80, 100) under a fixed number of facets (5); (2) varying the number of facets (5, 10) under a fixed number of topics (50). The results under fixed number of facets and varying number of facets are shown in Tables 7 and 8 respectively.

As we can see in Table 7, the full model (EFTMWE-MS) performs best on Dataset 1 and 2 when the number of topics is set to 30 and 80, respectively. Dataset 2 is a much bigger dataset compared to Dataset 1 and 30 topics appears to be enough to capture the topical patterns of the smaller dataset but is insufficient for the bigger dataset. In terms of the number of facets, we can see in Table 8 that the performance with just five facets is consistently better than the performance with 10 facets, which indicates that a small number of facets is probably enough to capture entity specific topics and a big number of facets might make things complicated. Overall, we can see that the performance varies much under different parameter settings, indicating a space of improvements by tuning parameters. We leave the tasks of finding the optimal number of facets and topics as future work.

Finding 4
A limited number of facets is enough to capture the different facets of an entity, on average.

9. Conclusions

In this work, we propose a model and an algorithm to learn entity-centric document representations, where a document associated with multiple entities is represented by multiple representations and each representation is built on the basis of entity facets. We demonstrate the effectiveness of our model, EFTM, by comparing it against state-of-the-art document representation methods on the task of multi-label classification, confirming that multi-faceted entity-centric modeling of documents has an effect in downstream applications. Although we evaluate our method on multi-label classification, our method is more broadly applicable to other multi-labeled settings where documents are associated with multiple source entities, such as tag analysis (Li, Guo, & Zhao, 2008) and tag suggestion (Katakis, Tsoumakas, & Vlahavas, 2008). We further investigated the notion of facets we learn by performing both an intrinsic and an extrinsic evaluation and confirmed that learned facets can capture semantically similar facets of different entities.

The theoretical implication of this work is that entities should not be considered and modeled as monolithic concepts, with a single representation for every document associated with them, but instead thought as multi-faceted concepts, with different facets discussed in different documents. Gaining a deeper understanding of what do these facets precisely represent, how many facets each specific entity has, and whether these facets can be mapped to explicit categories is left for future work that can enable interesting methods of analyzing and visualizing document collections from the perspective of an entity, but also analyzing how entities are presented in a document corpus. In practice, our work demonstrates that such a multi-faceted entity consideration can have an impact in downstream applications.

Table 7
Comparing the performance of document representation methods on the task of multi-label classification, using Dataset 1 and Dataset 2. The evaluation metric is multi-label accuracy (ml-Acc). For our models the number of facets is set to 5. The number of topics of our model and other models considered are 30, 50, 80, and 100, as indicated in row two.

Number of Topics		30	50	80	100
Dataset 1	EFTMWO-SS	0.0377	0.0378	0.0303	0.0302
	EFTMWO-MS	0.0455	**0.0453**	**0.0437**	**0.0422**
	EFTMWE-SS	0.0356	0.0381	0.0389	0.0362
	EFTMWE-MS	**0.0488**	0.0447	0.0402	0.0414
Dataset 2	EFTMWO-SS	0.0641	0.0681	0.0713	0.0568
	EFTMWO-MS	**0.0779**	0.0737	0.0695	0.0659
	EFTMWE-SS	0.0698	**0.0767**	**0.0806**	**0.0781**
	EFTMWE-MS	0.0728	0.0717	0.0736	0.0661

Table 8
Comparing the performance of document representation methods on the task of multi-label classification, using Dataset 1 and Dataset 2. The values 5 and 10 in columns 3–6 indicate the number of facets.

		Micro F1		Macro F1	
		5	10	5	10
Dataset 1	EFTMWO-SS	0.0710	0.0679	0.0702	0.0677
	EFTMWO-MS	**0.0857**	0.0661	**0.0885**	**0.0718**
	EFTMWE-SS	0.0722	**0.0680**	0.0740	0.0685
	EFTMWE-MS	0.0843	0.0618	0.0836	0.0657
Dataset 2	EFTMWO-SS	0.1217	0.1109	0.1182	0.1141
	EFTMWO-MS	0.1361	0.1127	0.1331	0.1123
	EFTMWE-SS	**0.1395**	**0.1393**	**0.1366**	**0.1386**
	EFTMWE-MS	0.1323	0.1193	0.1332	0.1173

CRediT authorship contribution statement

Chuan Wu: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing - original draft. **Evangelos Kanoulas:** Resources, Writing - review & editing, Formal analysis, Supervision, Project administration, Funding acquisition. **Maarten de Rijke:** Resources, Writing - review & editing, Supervision, Project administration, Funding acquisition.

Acknowledgements

This research was supported by Ahold Delhaize, the Association of Universities in the Netherlands (VSNU), the Google Faculty Research Awards program, the NWO Innovational Research Incentives Scheme Vidi (016.Vidi.189.039), the NWO Smart Culture - Big Data / Digital Humanities (314-99-301), the H2020-EU.3.4. - SOCIETAL CHALLENGES - Smart, Green And Integrated Transport (814961), Chinese Scholarship Council, and the Innovation Center for Artificial Intelligence (ICAI). All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

References

Alam, M. H., Ryu, W.-J., & Lee, S. (2016). Joint multi-grain topic sentiment: Modeling semantic aspects for online reviews. *Information Sciences, 339*, 206–223.
Balog, K. (2018). *Entity-Oriented search.* Springer.
Bast, H., Buchhold, B., & Haussmann, E. (2016). Semantic search on text and knowledge bases. *Foundations and Trends in Information Retrieval, 10*(2–3), 119–271.
Bicalho, P., Pita, M., Pedrosa, G., Lacerda, A., & Pappa, G. L. (2017). A general framework to expand short text for topic modeling. *Information Sciences, 393*, 66–81.
Bielza, C., Li, G., & Larranaga, P. (2011). Multi-dimensional classification with Bayesian networks. *International Journal of Approximate Reasoning, 52*(6), 705–727.
Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.
Bouadjenek, M. R., Hacid, H., Bouzeghoub, M., & Vakali, A. (2016). Persador: Personalized social document representation for improving web search. *Information Sciences, 369*, 614–633.
Chang, J., Boyd-Graber, J., & Blei, D. M. (2009). *Connections between the lines: augmenting social networks with text. Kdd.* ACM169–178.
Chen, M. (2017). Efficient vector representation for documents through corruption. *arXiv preprint arXiv:1707.02377*.
Croft, W. B. (1981). Document representation in probabilistic models of information retrieval. *Journal of American Society for Information Science, 32*(6), 451–457.
Dai, H., Tang, S., Wu, F., & Zhuang, Y. (2018). Entity mention aware document representation. *Information Sciences, 430*, 216–227.
Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
Dragoni, M., Federici, M., & Rexha, A. (2018). An unsupervised aspect extraction strategy for monitoring real-time reviews stream. *Information Processing & Management*.
Dunietz, J., & Gillick, D. (2014). *A new entity salience task with millions of training examples. Eacl.* ACL205–209.
Erosheva, E., Fienberg, S., & Lafferty, J. (2004). Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences, 101*(suppl 1), 5220–5227.
Fetahu, B., Markert, K., & Anand, A. (2015). *Automated news suggestions for populating wikipedia entity pages. Cikm.* ACM323–332.
Griffiths, D., & Tenenbaum, M. (2004). Hierarchical topic models and the nested chinese restaurant process. *NIPS, 16*, 17.
Han, X., & Sun, L. (2012). *An entity-topic model for entity linking. Emnlp.* ACL105–115.
Hu, L., Li, J., Zhang, J., & Shao, C. (2015). *o-hetm: An online hierarchical entity topic model for news streams. Pakdd.* Springer696–707.
Katakis, I., Tsoumakas, G., & Vlahavas, I. (2008). Multilabel text classification for automated tag suggestion. *ECML PKDD Discovery Challenge,* 75–83.
Kazai, G., Yusof, I., & Clarke, D. (2016). *Personalised news and blog recommendations based on user location, facebook and twitter user profiling. Sigir.* ACM1129–1132.
Kim, H., Sun, Y., Hockenmaier, J., & Han, J. (2012). *Etm: Entity topic models for mining documents associated with entities. Icdm.* IEEE349–358.
Kraft, P., Jain, H., & Rush, A. M. (2016). *An embedding model for predicting roll-call votes. Emnlp*2066–2070.
Le, Q. V., & Mikolov, T. (2014). *Distributed representations of sentences and documents. Icml*1188–1196.
Li, C., Xing, J., Sun, A., & Ma, Z. (2016). *Effective document labeling with very few seed words: A topic model approach. Cikm.* ACM85–94.
Li, P., Wang, Y., Gao, W., & Jiang, J. (2011). *Generating aspect-oriented multi-document summarization with event-aspect model. Emnlp.* ACL1137–1146.
Li, P., Wang, Y., & Jiang, J. (2013). Automatically building templates for entity summary construction. *Information Processing & Management, 49*(1), 330–340.
Li, X., Guo, L., & Zhao, Y. E. (2008). *Tag-based social interest discovery. Www.* ACM675–684.
Li, X., Ouyang, J., & Zhou, X. (Ouyang, Zhou, 2015a). Centroid prior topic model for multi-label classification. *Pattern Recognition Letters, 62*, 8–13.
Li, X., Ouyang, J., & Zhou, X. (Ouyang, Zhou, 2015b). Supervised topic models for multi-label classification. *Neurocomputing, 149*, 811–819.
Li, X., Wang, Y., Zhang, A., Li, C., Chi, J., & Ouyang, J. (2018). Filtering out the noise in short text topic modeling. *Information Sciences, 456*, 83–96.
Liu, M., Fang, Y., Choulos, A. G., Park, D. H., & Hu, X. (2017). Product review summarization through question retrieval and diversification. *Information Retrieval Journal, 20*(6), 575–605.
Ma, B., Zhang, N., Liu, G., Li, L., & Yuan, H. (2016). Semantic search for public opinions on urban affairs: A probabilistic topic modeling-based approach. *Information Processing & Management, 52*(3), 430–445.

Mcauliffe, J. D., & Blei, D. M. (2008). *Supervised topic models. Nips*. Curran Associates, Inc.121–128.
McCallum, A. (1999). *Multi-label text classification with a mixture model trained by EM. Aaai workshop on text learning*. AAAI1–7.
Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). *Distributed representations of words and phrases and their compositionality. Nips*. Curran Associates Inc.3111–3119.
Nanni, F., Ponzetto, S. P., & Dietz, L. (2018). *Entity-aspect linking: Providing fine-grained semantics of entities in context. Jcdl*. ACM49–58.
Newman, D., Chemudugunta, C., & Smyth, P. (2006). *Statistical entity-topic models. Kdd*. ACM680–686.
Pennington, J., Socher, R., & Manning, C. (2014). *Glove: Global vectors for word representation. Emnlp*1532–1543.
Perotte, A. J., Wood, F., Elhadad, N., & Bartlett, N. (2011). *Hierarchically supervised latent dirichlet allocation. Nips*2609–2617.
Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.
Piccardi, T., Catasta, M., Zia, L., & West, R. (2018). *Structuring wikipedia articles with section recommendations. Sigir*. ACM665–674.
Qiu, Z., & Shen, H. (2017). User clustering in a dynamic social network topic model for short text streams. *Information Sciences, 414*, 102–116.
Ramage, D., Hall, D., Nallapati, R., & Manning, C. D. (2009). *Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. Emnlp*. ACL248–256.
Rao, J., & Wu, C. (2010). Pseudo–empirical likelihood inference for multiple frame surveys. *Journal of the American Statistical Association, 105*(492), 1494–1503.
Raviv, H., Kurland, O., & Carmel, D. (2016). *Document retrieval using entity-based language models. Sigir*. ACM65–74.
Reinanda, R., Meij, E., & de Rijke, M. (2015). *Mining, ranking and recommending entity aspects. Sigir*. ACM263–272.
Reinanda, R., Meij, E., & de Rijke, M. (2016). *Document filtering for long-tail entities. Cikm*. ACM771–780.
Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004). *The author-topic model for authors and documents. Uai*. AUAI Press487–494.
Salton, G., Yang, C.-S., & Yu, C. T. (1975). A theory of term importance in automatic text analysis. *Journal of American society for Information Science, 26*(1), 33–44.
Sandhaus, E. (2008). The New York Times annotated corpus. LDC, Philadelphia https://catalog.ldc.upenn.edu/ldc2008t19.
Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys, 34*(1), 1–47.
Sikchi, A., Goyal, P., & Datta, S. (2016). *Peq: An explainable, specification-based, aspect-oriented product comparator for e-commerce. Cikm*. ACM2029–2032.
Spina, D., Meij, E., de Rijke, M., Oghina, A., Bui, M. T., & Breuss, M. (2012). *Identifying entity aspects in microblog posts. Sigir*. ACM1089–1090.
Titov, I., & McDonald, R. (2008). *Modeling online reviews with multi-grain topic models. Www*. ACM111–120.
Tsoumakas, G., & Katakis, I. (2007). Multi-label classification: An overview. *International Journal of Data Warehousing and Mining, 3*(3), 1–13.
Van Gysel, C., de Rijke, M., & Kanoulas, E. (de Rijke, Kanoulas, 2016a). *Learning latent vector spaces for product search. Cikm*. ACM165–174.
Van Gysel, C., de Rijke, M., & Kanoulas, E. (2018). Neural vector spaces for unsupervised information retrieval. *ACM Transactions on Information Systems, 36*(4).
Van Gysel, C., de Rijke, M., & Worring, M. (de Rijke, Worring, 2016b). *Unsupervised, efficient and semantic expertise retrieval. Www*. ACM1069–1079.
Xiao, D., Ji, Y., Li, Y., Zhuang, F., & Shi, C. (2018). Coupled matrix factorization and topic modeling for aspect mining. *Information Processing & Management, 54*(6), 861–873.
Xiong, C., Callan, J., & Liu, T.-Y. (2017). *Word-entity duet representations for document ranking. Sigir*. ACM763–772.
Yu, J., Zha, Z.-J., Wang, M., & Chua, T.-S. (2011). *Aspect ranking: Identifying important product aspects from online consumer reviews. Hlt*. ACL1496–1505.
Yu, Q., & Lam, W. (2018). *Review-aware answer prediction for product-related questions incorporating aspects. Wsdm*. ACM691–699.
Zhang, Y., Mao, W., & Zeng, D. (2016). *A non-parametric topic model for short texts incorporating word coherence knowledge. Cikm*. ACM2017–2020.