Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2020) 28: 2244 – 2260

doi:10.3906/elk-1912-62

NET-LDA: a novel topic modeling method based on semantic document similarity

Ekin EKİNCİ[1*], **Sevinç İLHAN OMURCA**[2]

[1]Software Engineering Department, Faculty of Engineering, Doğuş University, İstanbul, Turkey
[2]Computer Engineering Department, Faculty of Engineering, Kocaeli University, Kocaeli, Turkey

Received: 09.12.2019 • **Accepted/Published Online:** 18.04.2020 • **Final Version:** 29.07.2020

Abstract: Topic models, such as latent Dirichlet allocation (LDA), allow us to categorize each document based on the topics. It builds a document as a mixture of topics and a topic is modeled as a probability distribution over words. However, the key drawback of the traditional topic model is that it cannot handle the semantic knowledge hidden in the documents. Therefore, semantically related, coherent and meaningful topics cannot be obtained. However, semantic inference plays a significant role in topic modeling as well as in other text mining tasks. In this paper, in order to tackle this problem, a novel NET-LDA model is proposed. In NET-LDA, semantically similar documents are merged to bring all semantically related words together and the obtained semantic similarity knowledge is incorporated into the model with a new adaptive semantic parameter. The motivation of the study is to reveal the impact of semantic knowledge in the topic model researches. Therefore, in a given corpus, different documents may contain different words but may speak about the same topic. For such documents to be correctly identified, the feature space of the documents must be elaborated with more powerful features. In order to accomplish this goal, the semantic space of documents is constructed with concepts and named entities. Two datasets in the English and Turkish languages and 12 different domains have been evaluated to show the independence of the model from both language and domain. The proposed NET-LDA, compared to the baselines, outperforms in terms of topic coherence, F-measure, and qualitative evaluation.

Key words: Aspect extraction, cooccurence relation, latent Dirichlet allocation (LDA), semantic similarity, topic modeling

1. Introduction

Nowadays, customers share what they like and what not and what is in their mind by clicking on a website from where they are. Thus online review websites have revolutionized the daily economic life by changing the way of companies by which they can boost their sales and marketing. Based on the survey conducted by BrightLocal in 2017 [1], 85% of the customers trust online reviews as much as personal recommendations. On this basis, the huge amount of electronic word-of-mouth data provides a goldmine to the data scientists for the evaluation of others' opinions about what they have purchased or benefited from. Scientists mine web media for sentiment analysis and evaluate costumers' choices and can also help others in the decision making process. Sentiment analysis is the process of learning people's opinions which are expressed about individuals, events, or properties. The major problem in sentiment analysis is extracting product aspects. An aspect can be defined as an attribute of an entity, for example, burger, salad, waiter of a restaurant, etc. In the aspect based sentiment analysis,

*Correspondence: eekinci@dogus.edu.tr

[1]BrightLocal (2017). Local Consumer Review Survey 2017 [online]. Website https://www.brightlocal.com/research/local-consumer-review-survey/ [accessed 02 January 2018].

the extraction of a sentiment defines an aspect and fine-grained information about the specific sentiment being proposed. Many studies have been carried out for the aspect extraction task and the most promising results are obtained by using topic models.

Topic models [1–6] are based on the simple idea that documents are mixtures of topics, where a topic is a probability distribution over words. The primary objective of topic models is to discover main topics of a corpus of documents that are expected to be thematically similar, cohesive and self-contained. Probabilistic topic modeling [7, 8] is widely used to discover the latent thematic structure of a document collection and represents a powerful method in machine learning.

LDA which is proposed by Blei et al. [5] and widely used in text mining, computer vision, and computational biology is one of the most successful probabilistic topic modeling algorithms. LDA is described as a generative probabilistic model of a corpus. The intuition behind LDA is that it models documents as a random mixture of latent topics, where each topic is a distribution of a particular set of words [4]. Even if it is said that topics in latent space are semantically coherent, LDA considers only word cooccurrence distribution in the corpus of documents and not the semantic knowledge contained therein [9]. Therefore, semantically related, coherent, and meaningful topics cannot be achieved by using only topic models because of limited capability. This can be considered as the main drawback of LDA and to deal with this drawback a few studies that use semantic knowledge have been presented [3, 10–15].

However, in the above studies [3, 10–15], semantic knowledge was incorporated into the model to extract semantically related topics while semantic similarity of documents was not taken into account. In the current work, the semantic similarity of documents is incorporated into the LDA as an adaptive semantic parameter of the model. This parameter is also considered as a base semantic measure for LDA and used to influence the document-topic distribution in LDA. For this purpose, a similarity network with documents as nodes and similarity degrees as the weight of edges is constructed to represent the semantic similarity of documents. Since the proposed LDA model considers a combination of semantically connected documents, we call the model NET-LDA. In the network, semantically similar documents are merged to compose a single document, and the semantically related words come together in this new document. Thus, the word cooccurrence that is the basic assumption behind LDA is strengthened semantically.

In this study, the concepts and named entities of documents are used to determine the similarity between the documents. The concepts and named entities are extracted by using a publicly available concept network Babelfy. Babelfy was created by the Department of Computer Science Linguistic Computing Laboratory, the Sapienza University of Rome [16]. A concept is defined as the smallest semantic unit, having a unique meaning. For instance, in the review sentence "The baguettes in this restaurant are very delicious", concepts of the baguette are "French" and "loaf". Named entities are defined as the names of real world objects. In the review sentence "The restaurant states in Washington Boulevard.", the named entity of Washington Boulevard is a "downtown Detroit, Michigan.". As it is seen, concepts and named entities can be used as the best indicators of semantic meaning. The determination of the semantic relatedness of documents, using only these semantic units instead of using all words in the documents, provides results that are both accurate and faster.

The contributions of this study are as follows: (i) When the proposed LDA models in the literature are examined in detail, we observe that the advantages of the topic model and semantic similarity of documents are first combined into the NET-LDA to obtain semantically related and coherent topics. (ii) Similar studies in the literature incorporate the semantic information in their model to enhance the document space semantically. However, in NET-LDA, the semantically similar documents are merged then similarity information is incor-

porated into the model as a new adaptive semantic parameter. (iii) Contrary to the existing LDA models, in the NET-LDA, multiword aspects, as well as word aspects, are extracted from documents. (iv) The NET-LDA model is language-independent. It is applicable to 284 different languages which are provided by Babelfy. To prove language independence, English and Turkish datasets are used for the experimental purpose. (v) The NET-LDA model is domain-independent. To demonstrate this, the user reviews in 12 different domains are used to extract aspects, which can be accepted as the main topics of the reviews. As a result, valid aspects are extracted for all domains. (vi) We evaluated the proposed model qualitatively and quantitatively and conducted experiments for LDA, LTM (lifelong topic model), AMC (topic modeling with automatically generated must-links and cannot-links) and NET-LDA. For quantitative analysis, the results were evaluated according to topic coherence and F-measure. For the qualitative analysis, we considered the aspects generated by all models in terms of semantic relation and ability to capture details. The evaluation results demonstrated the strength of the proposed NET-LDA and showed that it outperforms.

The rest of the paper is organized as follows: Section 2 introduces the literature review. Section 3 details NET-LDA and its realization steps. Section 4 describes the datasets that are used in the experiments and evaluation of the results of the NET-LDA in contrast with the baselines for the purpose of aspect extraction. Section 5 concludes the study.

2. Literature review

LDA was presented in 2003 and after that, some baseline approaches were presented as extensions of LDA. Dynamic topic model (DTM) is 1 of them with an ability to capture the evolution of topics in a sequentially organized corpus of documents [17]. A new concept-topic model that was proposed by Chemudugunta et al. [12] combined human defined ontological concepts and data-driven concepts together.

Another baseline approach is labeled LDA (LLDA), which incorporates supervision into LDA by training multiply labeled document models, where each document is taken as a mixture of topics and then it extracts each word from a topic [18].

Other 2 baseline approaches, namely, MedLDA and RTM, which are the extensions of LDA, were presented in the same year [19]. Zhu et al. [20] combined the intuition of the max-margin prediction models (such as SVM) with the intuition of hierarchical Bayesian topic models (such as LDA) and called their novel supervised topic model as the maximum entropy discrimination latent Dirichlet allocation (MedLDA) model. They conducted some experiments on movie reviews, hotel reviews, and 20 newsgroups datasets and extracted more discriminative latent topics. The relational topic models (RTM), which is another extension of LDA, is modeled as a network of documents and a link between them. It is used to analyze linked corpora such as citation networks [21].

Zhai et al. [22] presented the first constrained LDA model which uses must-links and cannot-links constraints to improve the accuracy of LDA for grouping the product features. While a must-link constraint specifies that 2 product features must be in the same cluster; a cannot-link constraint specifies that 2 product features cannot be in the same cluster. A novel parallelized LDA algorithm (Mr. LDA) in the MapReduce programming framework was proposed by Zhai et al. [23]. While the previous LDA approaches use Gibbs sampling, Mr. LDA relies on variational inference.

A novel approach for calibrating LDA on a text corpus about software engineering was proposed by Panichella et al. [24]. The authors used genetic algorithms to determine a near optimal configuration for LDA by considering 3 software engineering tasks such as traceability link recovery, feature location, and software

artifact labeling.

There are several studies that have been focused on LDA solutions for aspect extraction and opinion mining. Bagheri et al. [25] presented a novel approach based on LDA for aspect extraction problem in sentiment analysis systems. Their approach assumes that aspects in a sentence form a Markov chain instead of a bag-of-words assumption. Another work on identifying aspects and sentiment words from reviews is called as appraisal expression patterns based latent Dirichlet allocation (AEP-LDA) and presented by Zheng et al. [26]. The authors incorporated appraisal expression patterns into LDA for aspect and sentiment word identification. All words in a single sentence are drawn from one topic and every corpus is composed of corresponding aspect and sentiment topics.

Wang et al. [27] presented 2 semisupervised topic models for product aspect extraction: the fine-grained labeled LDA (FL-LDA) and the unified fine-grained labeled LDA (UFL-LDA). At first, the seeding aspects and related seeding words are extracted from detailed product reviews, and then these seed sets are used in the proposed topic models.

Xie et al. [28] proposed a Markov random field (MRF) regularized LDA model and incorporated word similarities into topic modeling. They proposed an MRF for the latent topic layer of LDA to extract similar words to be put into the same topic. This can extract a word similar to other words under different topics.

LDA model is combined with the Girvan-Newman algorithm, which is a betweenness-based community detection algorithm by Li et al. [29]. The authors aimed to provide detailed and impressive information about detected communities. Liu et al. [30] presented a personality trait LDA model, which defines the personality recognition problem based on a topic model. Their proposed model was aimed to predict the personality traits and mine user behaviors such as topic preferences. Topics are constituted by the multinomial distribution over words and Gaussian distributions over personality traits. Zoghbi et al. [31] proposed multi-idiomatic LDA (MiLDA), which links related information from different sources of the same language for suggesting a solution to intrinsic multi-idiomatic data.

In another study [32], a novel conceptual dynamic latent Dirichlet allocation model (CDLDA) for topic detection on conversational content was presented, which employed a concept-based semantic representation of words. The presented model can extract the dependencies between topics and speech acts. Ekinci and İlhan Omurca [33] devised concept-LDA to obtain semantically related topics from user reviews in English. For this purposed they enriched reviews with semantic units extracted with Babelfy.

Rao [34] proposed a multilabeled sentiment topic model for adaptive social emotion classification, which was called contextual sentiment topic model (CSTM). By this model, they aimed to distinguish explicitly generalized topics and the topics that characterize context-dependent information across different collections.

Xie et al. [35] proposed a novel sketch-based topic model (TopicSketch), which detects twitter topics in real time. Their framework can detect different topics in a stream of more than 30 million tweets. For this challenging problem, instead of keeping a set of active words, they proposed a novel solution by hashing the active words into a small number of buckets due to word size. Since extracting semantic aspects across different domains is challenging, Alam et al. [36] proposed a domain-independent topic sentiment model, called joint multi-grain topic sentiment (JMTS), to extract semantic aspects. JMTS model was tested for online reviews and could extract sentiment-oriented aspects automatically by eliminating manual probe. In the JMTS model, domain-independent sentiment words are used as the prior sentiment information and then these priors are exploited by the model.

Zhang et al. [15] devised an idea discovery graph-based LDA-IG which combines cooccurrence and semantic relations of words with LDA. They, by employing the proposed model, could effectively detect rare important topics.

Yao et al. [37] aimed to improve previous knowledge based topic models by proposing a word embedding LDA (WE-LDA) model, which combines Skip-Gram and LDA. They proved the effectiveness of their model by analyzing real world e-commerce datasets. Fu et al. [38] proposed a topic sentiment joint model with word embeddings. They improved the topic identification and sentiment recognition by incorporating word embeddings and HowNet lexicon into topic modeling.

Sham and Baraani-Dastjerdi [39] presented ELDA by incorporating cooccurrence relations as prior domain knowledge into the LDA to extract more precise aspects from English and Persian datasets.

Heng et al. [40] applied the LDA to discover the underlain topics from online food and grocery shopping reviews.

Kandemir et al. [41] proposed a new way to supervise LDA with Gaussian processes as nonlinear predictors. Thus they can perform topic modeling and document classification jointly.

3. NET-LDA model

The proposed model contains 2 main modules: semantic network construction (NET) and topic extraction with an incorporated topic-document distribution parameter (LDA) as illustrated in Figure 1. The first module includes 5 steps: (I) preprocessing, (II) multiword term extraction, (III) concept and named-entity extraction, (IV) (a) calculating document similarity through concepts and (b) named entities and drawing graph between similar documents, (V) merging similar documents.

In our approach, various corpora from different domains and languages are fed separately into NET-LDA as the input and a set of aspects under topics for each domain is obtained as the output. Herein, with the term "domain", we intend to define the set of documents that refer to a specific product, person, or service. This set of documents is denoted by $D = \{d_1, d_2, , d_M\}$. In order to achieve a better aspect representation, multiword aspects are taken into account. Thus by using Babelfy, all multiwords from the documents are extracted in step (I), while in step (II) basic preprocessing steps are applied to each document. Further, in step (III), concepts and named entities are extracted from every d_j in D to strengthen cooccurrence relation with semantic knowledge. For each document of the corpus, d_j, a new document, d'_j, which contains only concepts and named-entities, is created. This new corpus, which is represented semantically is denoted by $D' = \{d'_1, d'_2, , d'_M\}$. In step (IV), semantic similarity between d'_j's of D' is calculated by using cosine similarity. By using this similarity, we draw a similarity network between d_j's. In step (V), all similar documents in D are merged by using a similarity network of the related documents based on D' to create a set of new documents $MD = \{md_1, md_2, , md_G\}$. For instance, as shown in Figure 1, if d'_1, d'_3, d'_5 and d'_M are similar documents, then these documents are merged and in this case the value of u_i is 4. Finally, topics are extracted from these merged documents.

In order to explain steps (III), (IV), and (V) of NET-LDA, 3 user reviews from computer domain are selected and given in Figure 2. The first one is d_1 = "power supply died and laptop will not power on, screen is very small", the second one is d_2 = "I absolutely love this notebook the best display I have reviewed", and the last one is d_3 = "the mouse pad is covered with a high quality fabric". The words that contain "_" represent multiwords in the reviews. When these review sentences are examined, it can be seen that d_1 and d_2 contain synonyms. In these sentences, "laptop" and "notebook" and "screen" and "display" are synonyms

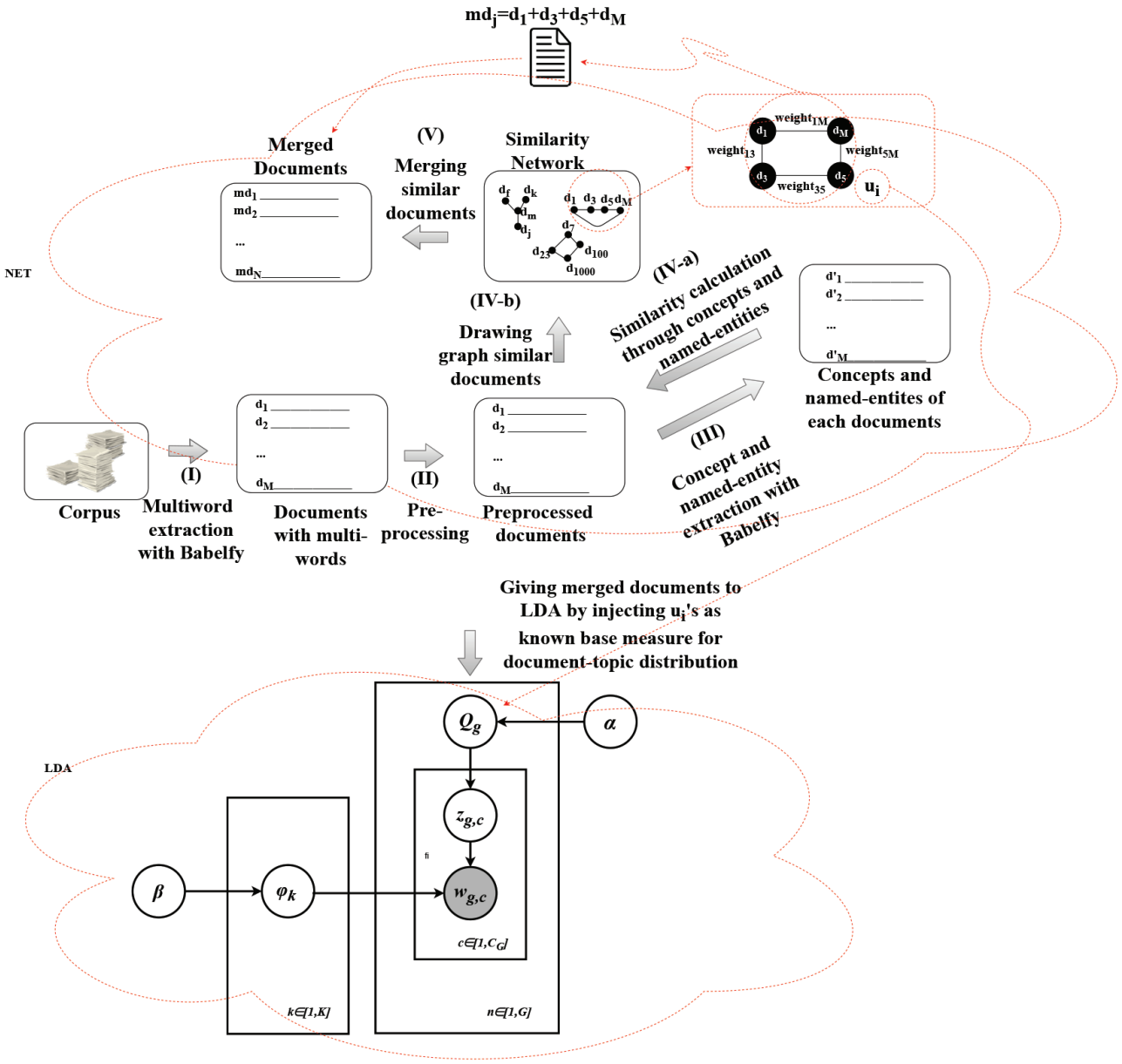

Figure 1. Steps of NET-LDA.

within themselves. However, the similarity between these 2 sentences is zero and these synonyms cannot be placed in the same topic. Therefore, in step (III), the concepts and named entities are extracted from the reviews independently. By extracting concepts and named-entities we obtain d'_1 = "electric power electrical load portable computer lap screen display computer user" from d_1, d'_2 = "qualification portable computer lap good screen display computer user" from d_2 and d'_3 = "portable ball computer mouse artifact weaving knitting crocheting fibers" from d_3. If user reviews are examined on the basis of concepts and named entities, then synonyms can easily be identified. In step (IV-a), the similarity between the concepts and named entities of the reviews is calculated. In this case, while similarity between d_1 and d_2 is zero, the similarity between d'_1 and d'_2

is equal to 0.537. Since the similarity is greater than 0.5, the connection between d_1 and d_2 is permitted as in step (IV-b). The similarity between d'_1 and d'_3 and that between d'_2 and d'_3 is also zero, thus the connections between d_1 and d_3 and d_2 and d_3 are prohibited in step (IV-b). After step (IV-b) is realized, only 1 edge is constructed in the network and this edge is between d_1 and d_2. Consequently, md_1 is composed by merging d_1, d_2, and d_3 as a single node graph and referred to as md_2. At the beginning of step (III), there are 3 documents in D, at the end of step (V), there are only 2 documents in MD. Since md_1 contains 2 documents, the value of adaptive semantic parameter u_1 is 2 for md_1. In a similar manner, as md_2 contains only 1 document, the value of adaptive semantic parameter u_2 is 1 for md_2. Briefly, after these 3 steps, the set of merged documents and adaptive semantic parameters u's are obtained. Each element of MD and the semantic adaptive semantic parameter u are given as the inputs of LDA to extract topics.

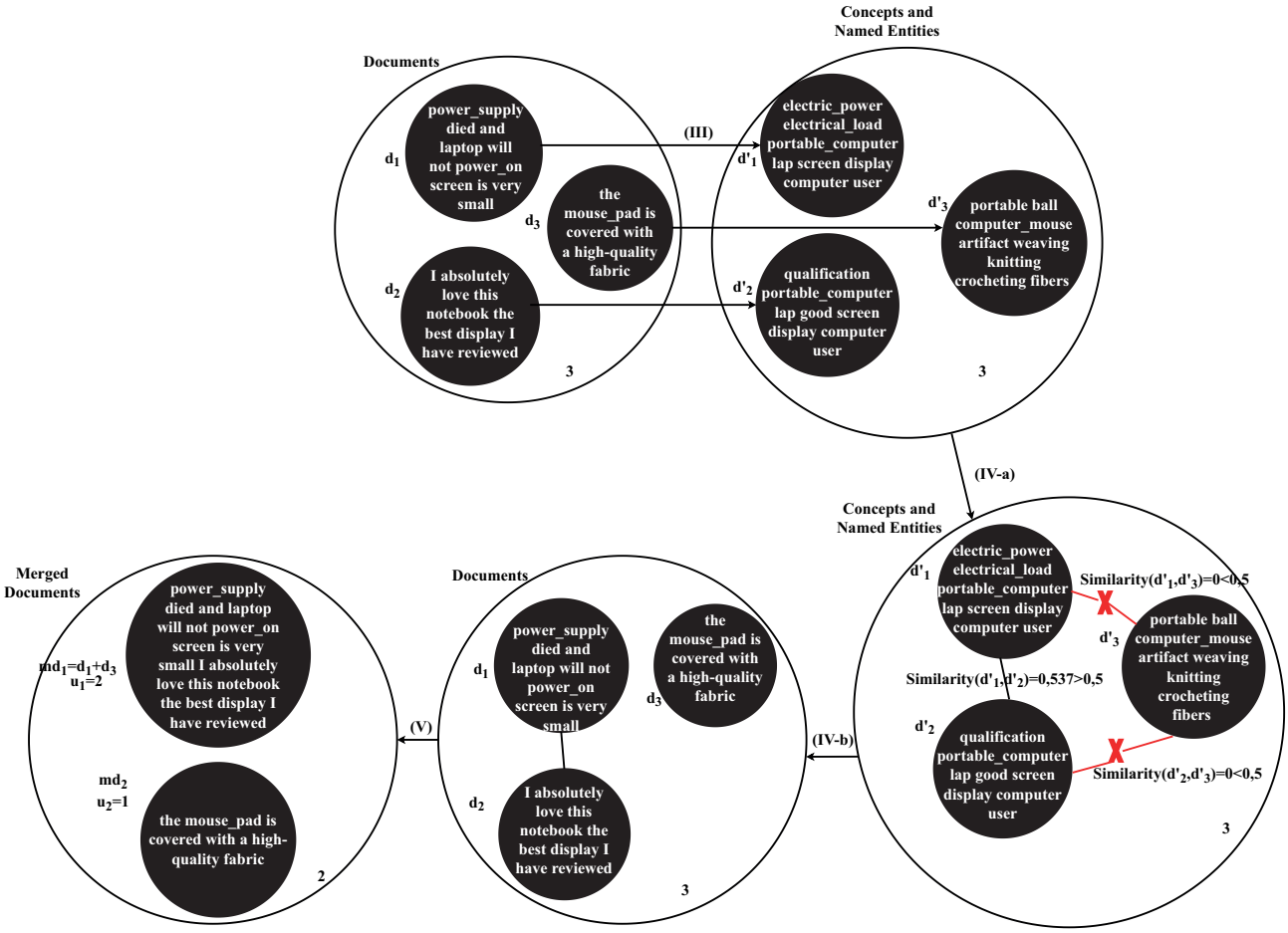

Figure 2. The detailed steps of the NET-LDA.

3.1. Multiword extraction

In the review sentences, product aspects are available in multiword expressions as well as word expressions, such as "hard drive", "refresh rat", and "Android market". In order to obtain fine grained aspects, multiword aspect extraction is required. Therefore, multiword expressions in the documents are extracted by using Babelfy and all of its properties. The extracted multiword terms from the experimental datasets could be categorized into 3 types in the reviews: the multiwords taking place in dictionary (hard drive), the domain-based multiwords

(refresh rate), and the multiword named entity phrases (Android market).

3.2. Preprocessing

Preprocessing is a very crucial task in terms of managing and analyzing the data before it is fed into the model as an input. In this study, we used corpora from 2 different languages: English and Turkish, but performed the same preprocessing tasks for all corpora. We performed a spell correction then removed stop words, which are regarded as noisy labels, URLs, punctuations, and digits, which are not meaningful for our task. Then, all letters in the capital were converted to lowercase. In the stemming process, we used Zemberek Turkish natural language processing library [42] and Stanford natural language processing tool [43].

3.3. Concept and named-entity extraction

The main purpose of NET-LDA is to strengthen word cooccurrences semantically. Therefore, we represent document space semantically by using concepts and named entities as the semantic units instead of words. For this purpose, Babelfy is used to extract concepts and named-entities from documents. The reasons why we choose Babelfy for the concept and named entity extraction can be summarized as follows: (I) By using Babelfy, not only the meanings of documents are obtained but also word sense ambiguity is resolved. For example, in the review sentence what does the reviewer mean with "driver", "the operator of a motor vehicle" or "a program that determines how a computer will communicate with a peripheral device", (II) Babelfy covers 284 different languages, including languages spoken by several million people such as Sango and Tagalog, and languages spoken by billions of people such as English and Turkish so language-independence is ensured, (III) It merges 13 different resources. Thus it does not need to access any other source for information [44, 45].

3.4. Similarity network construction

In the document similarity graph construction phase, a set of concepts and named-entities D', which is extracted from document set D, is used. By using both concepts and named-entities accurate meaning of the words in documents are obtained and word sense ambiguity is resolved. Thus, by calculating document similarity based on concepts and named-entities, more accurate and effective similarity results can be obtained. The proposed algorithm is given as Algorithm 1.

In order to compute the similarity of documents through concepts and named-entities, firstly tf-idf term weighting scheme is used to represent each d'_j in D' then the cosine similarity measure is used to compute the pairwise similarities of d'_j's. For every d'_j most similar d'_t in D' is determined so that the similarity between d'_j and d'_t is greater than the user defined threshold and maximum similarity for d'_j. After using this pairwise similarity, we draw graph between d_j and d_t, the connection weights of edges indicate this similarity. In the constructed network, single node graphs, 2 node graphs, and also multinode graphs can be seen. All documents in each separate graph are merged. The total number of documents in a given corpus is equal to the total number of graphs in the constructed network; thus it means that the total number of documents is decreased. For each new document, the number of documents (node count in the graph) that are merged to comprise these new documents is held for NET-LDA model.

Input : $D' = \{d'_1, d'_2, ..., d'_M\}$ set of concepts and named entities of $D = \{d_1, d_2, ..., d_M\}$
Output: $DocumentSimilarityGraph(D)$, set of similar documents to D, a network with documents as nodes and graphs which connect similar documents $d_j \in D$ as nodes

```
foreach d'_j ∈ D' do
    represent d'_j with tf-idf weighting scheme
end
foreach d'_j ∈ D' do
    max(d'_j) = ∅
    maxSimilarity(d'_j) = −∞
    foreach d'_t ∈ D' do
        if j ≠ t then
            calculate cosine(d'_j, d'_t)
            if cosine(d'_j, d'_t) > thresholddegree and cosine(d'_j, d'_t) > maxSimilarity(d'_j) then
                max(d'_j) = d'_t
                maxSimilarity(d'_j) = cosine(d'_j, d'_t)
            end
        end
    end
end
foreach d_j ∈ D do
    if max(d'_j) ≠ ∅ then
        find max(d'_j) = d'_t
        draw graph between d_j and d_t
    end
end
```

Algorithm 1: Similarity network construction algorithm.

3.5. Topic extraction

In the literature, LDA is mostly developed by using symmetric Dirichlet priors due to the general opinion that these priors have little effect on the LDA. However, Wallach et al. found that using the asymmetric Dirichlet priors over the document-topic distributions has a great influence on the model performance [46]. By considering the effect of Dirichlet prior optimization on performance improvement and the fact that similar documents have similar topic distribution, asymmetric Dirichlet priors are applied in NET-LDA. We used asymmetric Dirichlet priors over the document-topic distributions because it offers much better model performance, measured in terms of the coherence and quality of extracted topics.

Asymmetric Dirichlet prior based NET-LDA is proposed as shown in Figure 3. In order to produce samples from posterior NET-LDA, we used Gibbs sampling. The sampling process is as follows: (i) words' distributions under topics are sampled by φ with a Dirichlet prior, (ii) for document-topic distribution, as illustrated by Wallach et al., we introduced node count in the graph. While sampling document-topic distributions, we used asymmetric Dirichlet priors over θ with a Dirichlet prior α and a known semantic measure u_g. Herein, u_g is incorporated into the model via edge from every topic assignment. The parameters of NET-LDA are given in Table 1.

According to the given model, for θ, we used asymmetric Dirichlet priors with hyperparameter α and a known semantic measure u. The proposed conditional posterior probability of topic k in document g given z

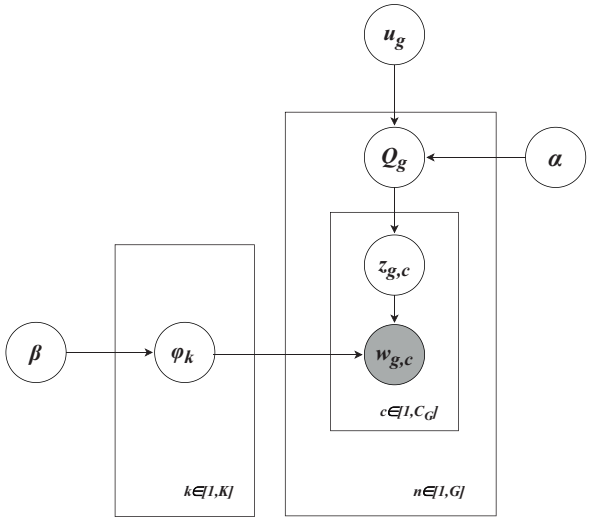

Figure 3. NET-LDA.

Table 1. Model parameters.

Parameter	Explanation
G	Number of documents in document set MD_i
K	Number of latent topics
C_g	Number of words in document g
α, β	Hyperparameters of NET-LDA
u_g	A known semantic measure for document g
θ_g	Multinomial distribution over topics for document g
φ_k	Multinomial distribution over words for topic k
z_g, c	Latent topic of word c in the g^th document
w_g, c	The c^th word in the g^th document

as:

$$p(z_{g,c} = k|z_{-g,c}, \alpha u) = \int d\theta_g p(k|\theta_g) p(\theta_g|z_{-g,c}, \alpha u) = \frac{c_{g,k} + \alpha u}{C_g - 1 + K\alpha u}. \tag{1}$$

$p(z_{g,c} = k|z_{-g,c}, \alpha u)$ is the probability of the current word c_g in document g assigned to each topic, conditioned other than all other words c_{-g}. If topic k is not included in document g, $c_{g,k}$ (number of words in document g which are assigned to topic k) will be zero. By using an asymmetric Dirichlet prior, we smooth c_g, k with αu. Since, the current word is not included, 1 is subtracted from C_g number of words in document g. Based on formula (1), full conditional distribution $p(z_{g,c} = k|z_{-g,c}, w)$ can be calculated as follows:

$$p(z_{g,c} = k|z_{-g,c}, w) = \frac{c_{g,k} + \alpha u}{C_g - 1 + K\alpha u} \frac{c_{w,k} + \beta}{\sum_{w' \in V} c_{w',k} + V\beta}. \tag{2}$$

In formula (2), with second ratio probability of topic k in document, g is expressed. $c_{w,k}$ defines the number of times word w is assigned to topic k in the whole collection. $c_{w',k}$ is the number of times topic k is used in the whole collection but not including the current word.

4. Numerical experiments

4.1. Datasets

We employed 2 different datasets to evaluate our proposed NET-LDA model. The first one is in Turkish and contains only 1 domain with 1517 reviews from a Turkish tourism website www.otelpuan.com [2]. The second one in English and contains 11 domains (Dataset1K) from www.amazon.com. More information about the datasets can be found in an earlier report [47]. In each domain, there are different numbers of reviews. If the similarity of the concepts of the 2 documents is greater than 0.5, we merge these 1 documents to compose a new document. The list of domains, number of reviews for original dataset, and similarity based dataset, number of words and multiwords in each domain are given in Table 2.

Table 2. Summary of domains.

#	Domain	Language	# Docs (original)	# Docs (similarity)	# Words	# Multiwords
1	Hotel	Turkish	1517	494	505	421
2	Alarm clock	English	5113	1268	694	385
3	Amplifier	English	5731	1341	923	712
4	Battery	English	4056	1004	549	251
5	Blu-ray player	English	9170	2052	1041	883
6	Cable modem	English	5754	1343	691	457
7	Camcorder	English	8361	1716	1049	898
8	Camera	English	8958	2112	1218	980
9	Car stereo	English	5587	1347	761	518
10	Cell phone	English	5713	1435	924	608
11	Computer	English	9090	1941	1251	1140
12	DVD player	English	6256	1560	871	647

4.2. Parameter settings

NET-LDA executed 50, 100, 200, 500, and 1000 iterations of Gibbs sampler. The number of topic that is denoted by K is set to 100. The Dirichlet hyperparameters α and β are assigned as $\frac{50}{K}$ and 0.001, respectively. For each new document, u is the number of merged documents used to compose the new document. Each extracted topic is identified by using first ten words. For comparison, same parameter settings are used for baselines.

4.3. Compared topic models

For comparison, we used LDA, LTM, and AMC topic models. LDA is the basic and simplest topic modeling as no prior knowledge is involved in this model [5]. LTM is a lifelong learning topic model, which uses must-links as prior knowledge to obtain coherent and better topics [47]. AMC is also a lifelong learning topic model and extracts must-links and cannot-links as prior knowledge to investigate better topics [48].

[2]Topic Extraction Dataset (2018). Turkish Hotel Reviews [online]. Website http://dx.doi.org/10.13140/RG.2.2.33914.80323 [accessed 20 June 2018].

4.4. Measures for performance evaluation

In order to measure semantic coherence of the topics extracted with NET-LDA and compare the NET-LDA with LDA, LTM, and AMC, we used topic coherence [49] and precision, recall and F-measure as evaluation metrics. Topic coherence is computed as follows:

$$C(k;V^k)=\sum_{n=2}^{N}\sum_{l=1}^{n-1}=\log\frac{D(v_n^{(k)},v_l^{(k)})+1}{D(v_l^{(k)})}. \tag{3}$$

In formula (3), $V^{(k)}=(v_1^{(k)},v_2^{(k)},,v_S^{(k)})$ is the list of most probable S words in topic k. $D(v_n^{(k)},v_l^{(k)})$ is the co-document frequency of word types v_n and v_l, and 1 is used for smoothing. $D(v_l^{(k)})$ is the document frequency of word type v_l. The higher value of the topic coherence reflects a high quality of the extracted words.

In order to evaluate the performance and the robustness of our approach on the basis of precision, recall, and F-measure, human-annotated topic words were obtained for each domain from the datasets. The precision of the topic model was calculated as the intersection between the relevant topic words, which are extracted by human-annotation and the whole topic words extracted by the topic model. The recall of the topic model is calculated as the intersection between the relevant topic words extracted by human-annotation and the relevant topic words extracted by the topic model. F-measure is calculated as the harmonic mean of precision and recall.

4.5. Experimental results

In this study, we developed an efficient topic model by writing a code in Java for extracting semantically related, coherent and meaningful topics and presented qualitatively and quantitatively performance success of NET-LDA over baseline models on 2 different datasets with 12 domains. The average topic coherence and F-measure for NET-LDA are determined by using datasets merged based on similarity and, for LDA, LTM, and AMC, they are given by using original datasets as shown in Figures 4 and 5, respectively.

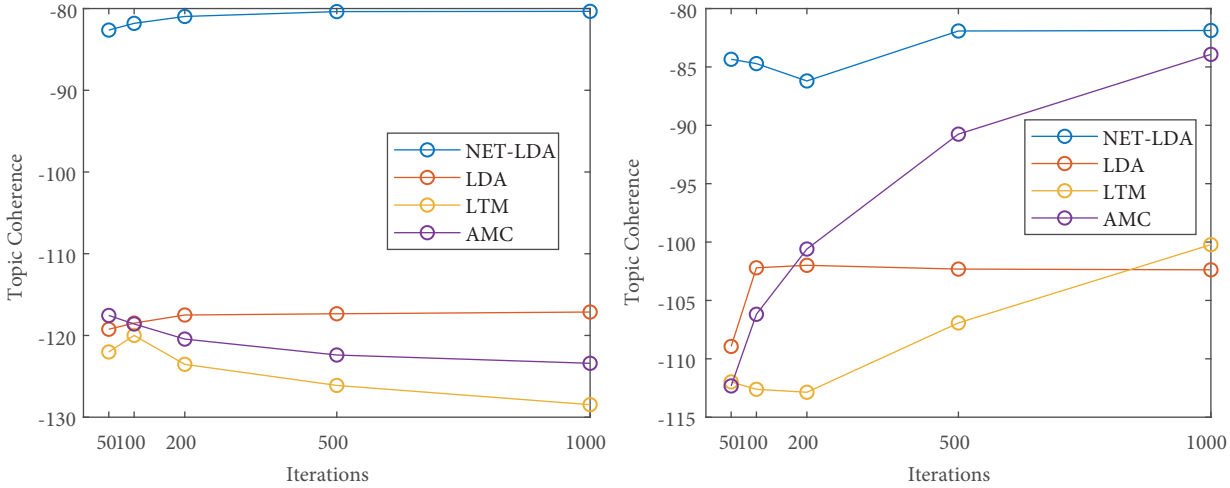

Figure 4. Topic coherence of each model on the English (left) and Turkish (right) datasets.

As can be seen in Figure 4: (I) NET-LDA performs better at all baselines in terms of topic coherence for Turkish and English datasets; this proves that using strengthened the cooccurrence relation is very effective.

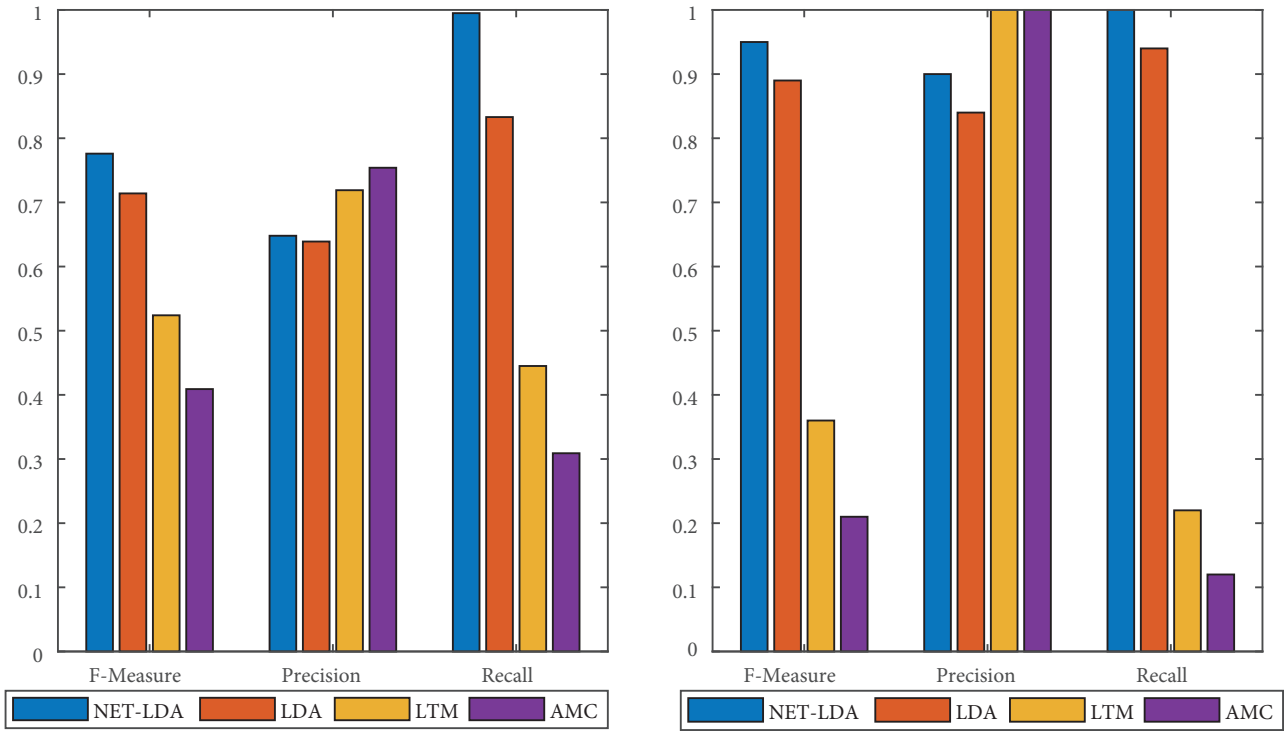

Figure 5. F-measure, precision and recall of each model on the English (left) and Turkish (right) datasets.

Thus, we can say that NET-LDA achieves high quality topics than the baselines. (II) For the English dataset, LTM and AMC showed poor performance because these models mine frequent item sets as prior knowledge shared among the topics, thus high-frequent words are contained in almost every topic. In Computer domain, the ”charge” can be given as an example to describe this situation. This may reduce the distinctiveness and quality of topics. (III) When the topic coherence for Turkish is examined, it is seen that LTM and LDA are the weakest. The use of general knowledge is the reason behind the low success of LTM. Using cannot-links improves the performance of the AMC for Turkish.

We give a detailed explanation for the improvements in NET-LDA by using F-measure. The improvement in NET-LDA is about 6% over LDA, 27% over LTM, 40% over AMC. As NET-LDA can extract more detailed topics, LTM and AMC can extract more general topics. By using more strengthened cooccurrence relation, NET-LDA can extract aspect words such as “battery life” and ”battery power” whereas LTM and AMC cannot extract these aspects as they do not include cooccurrence relation into the prior knowledge. Further, the intuition behind AMC and LTM provides better precision than NET-LDA, besides Compared to NET-LDA this intuition leads to worse recall for both of them. However, for more fine-grained aspect based sentiment analysis, aspect extraction should be performed in more detail.

For the qualitative analysis, top 10 topic words are presented in Table 3 as the most emphasized aspects of computer domain.

The extracted topics are labeled manually by using most likely words in the topics. The qualitative results show that the extracted topics of NET-LDA are semantically related, coherent and meaningful; however, the baseline results are of quite low quality. Therefore, we can easily label the topics extracted with NET-LDA. Furthermore, with NET-LDA, we can obtain detailed product aspects such as “dvi cable”, “dvi port” and

Table 3. Example topics obtained from LDA models in the computer domain. Errors are italicized in red.

Speaker				LCD			
NET-LDA	LDA	LTM	AMC	NET-LDA	LDA	LTM	AMC
speaker	speaker	speaker	speaker	lcd	lcd	lcd	lcd
sound	sound	sound	sound	dvi	*expensive*	monitor	crt
external	external	stereo	*monitor*	vga	*wrong*	crt	*thrive*
headphone	loud	*tinny*	*samsung*	dvi cable	hd	*mind*	picture
stereo	stereo	headphone	*feature*	input	*headphone*	toshiba	flat panel
ca	*blurry*	audio	*power*	lcd screen	acer	viewsonic	*expensive*
loud	*package*	*separate*	analog	*strong*	*process*	*side*	*software*
audio	*base*	*experience*	*button*	dvi port	view	*simple*	*nec*
headphone jack	*scratch*	*load*	*hp*	interface	*exchange*	*beautiful*	*version*
bell	notebook	*report*	digital	*optional*	viewing	*music*	light

"headphone jack". Consequently, when we compare the NET-LDA with baselines for the computer domain; it is obvious that the topic words of NET-LDA provide a superior representation of the computer domain.

5. Conclusions

Semantically related, coherent, and meaningful topics cannot be obtained by using only the LDA model because it considers only word cooccurrence distribution in the document corpus and not the semantic knowledge contained therein. This can be considered as the main drawback of LDA. In this paper, to deal with this drawback, a novel semantic network based LDA model is proposed. For this aim, a known semantic measure, which is obtained from a number of semantically similar documents, is introduced into the proposed model to influence document-topic distribution. In order to obtain semantically similar documents, feature space of documents is constructed by concepts and named entities instead of words that compose the documents. The representation of the feature space with these units ensures that the semantic knowledge is correctly included in the model. The experimental results of NET-LDA model indicate that injection of semantic similarity to the model provides semantically related, coherent and meaningful topics. NET-LDA does not depend on the domain and language of the corpora. While some frequent words appear in several final topics in LDA or in other baselines, NET-LDA can cope with the domination of these words. To the best of our knowledge, LDA has not been improved by using semantic document similarity earlier. Our experimental results on corpora indicate that NET-LDA achieves significant improvements over baselines.

References

[1] Hofmann T. Probabilistic Latent Semantic Analysis. In: Fifteenth Conference on Uncertainty in Artificial Intelligence; Stockholm, Sweden; 1999. pp. 289-296.

[2] Hofmann T. Unsupervised Learning by Probabilistic Latent Semantic Analysis. Machine Learning 2001; 42 (1-2): 177-196. doi: 10.1023/A:1007617005950

[3] Griffiths TL, Steyvers M. A probabilistic approach to semantic representation. In: Twenty-Fourth Annual Conference of the Cognitive Science Society; Fairfax, Virginia, USA; 2002. pp. 381-386.

[4] Griffiths TL, Steyvers M. Prediction and semantic association. In: Becker S, Thrun S, Obermayer K (editors). Advances in neural information processing systems. Cambridge, MA, USA: MIT Press, 2003, pp. 11-18.

[5] Blei DM, Ng AY, Jordan MI. Latent Dirichlet Allocation. Journal of Machine Learning Research 2003; 3: 993-1022.

[6] Griffiths TL, Steyvers M. Finding scientific topics. Proceedings of the National Academy of Sciences 2004; 101 (suppl 1): 5228-5235. doi: 10.1073/pnas.0307752101

[7] Steyvers M, Griffiths TL. Probabilistic topic models. In: Landauer TK, McNamara DS, Dennis S, Kintsch W (editors). Handbook of latent semantic analysis. Washington, DC, USA: Lawrence Erlbaum Associates Publishers, 2007, pp. 427-448.

[8] Blei DM. Probabilistic topic models. Communications of the ACM 2012; 55 (4): 77-84. doi: 10.1145/2133806.2133826

[9] Chang J, Gerrish S, Wang C, Blei DM. Reading tea leaves: How humans interpret topic models. In: Bengio Y, Schuurmans D, Lafferty JD, Williams CKI, Culotta A (editors). Advances in Neural Information Processing Systems 22. New York, NY, USA: Curran Associates Inc., 2009, pp. 288-296.

[10] Kim SM, Hovy E. Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text. In: Workshop on Sentiment and Subjectivity in Text; Sydney, Australia; 2006. pp. 1-8.

[11] Griffiths TL, Steyvers M, Tenenbaum J. Topics in semantic representation. Psychological Review 2007; 114 (2): 211-244. doi: 10.1037/0033-295X.114.2.211

[12] Chemudugunta C, Holloway A, Smyth P, Steyvers M. Modeling documents by combining semantic concepts with unsupervised statistical learning. In: Sheth A, Staab S, Paolucci M, Maynard D, Finin T et al. (editors). The Semantic Web - ISWC 2008. Heidelberg, Germany: Springer, 2008, pp. 229-244.

[13] Godin F, Slavkovikj V, De Neve W, Schrauwen B, Van de Walle R. Using topic models for twitter hashtag recommendation. In: 22nd International Conference on World Wide Web (WWW '13 Companion); Rio de Janeiro, Brazil; 2013. pp. 593-596.

[14] Poria S, Chaturvedi I, Cambria E, Bisio F. Sentic LDA: Improving on LDA with semantic similarity for aspect-based sentiment analysis. In: International Joint Conference on Neural Networks (IJCNN); Budapest, Hungary; 2016. pp. 4465–4473.

[15] Zhang C, Wanga H, Caoc L, Wanga W, Xu F. A hybrid term–term relations analysis approach for topic detection. Knowledge-Based Systems 2016; 93: 109-120. doi: 10.1016/j.knosys.2015.11.006

[16] Moro A, Raganato A, Navigli R. Entity linking meets word sense disambiguation: a unified approach. Transactions of the Association for Computational Linguistics 2014; 2: 231-244. doi: 10.1162/tacl_a_00179

[17] Blei DM, Lafferty JD. Dynamic Topic Models. In: 23rd International Conference on Machine Learning (ICML '06); Pittsburgh, Pennsylvania, USA; 2006. pp. 113-120.

[18] Ramage D, Hall D, Nallapati R, Manning CD. Labeled LDA: A Supervised Topic Model for Credit Attribution in Multi-labeled Corpora. In: 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP'09); Singapore; 2009. pp. 248-256.

[19] Jelodar H, Wang Y, Yuan C, Feng X. Latent Dirichlet allocation (LDA) and topic modeling - models, applications, a survey. Multimedia Tools and Applications 2019; 78 (11): 15169-15211. doi: 10.1007/s11042-018-6894-4

[20] Zhu J, Ahmed A, Xing EP. MedLDA: maximum margin supervised topic models for regression and classification. In: 26th Annual International Conference on Machine Learning (ICML '09); Montreal, Quebec, Canada; 2009. pp. 1257-1264.

[21] Chang J, Blei DM. Relational topic models for document networks. In: 12th International Conference on Artificial Intelligence and Statistics (AISTATS); Clearwater Beach, Florida, USA; 2009. pp. 81-88.

[22] Zhai Z, Liu B, Xu H, Jia P. Constrained LDA for grouping product features in opinion mining. In: Huang JZ, Cao L, Srivastava J (editor). Advances in Knowledge Discovery and Data Mining. Heidelberg, Germany: Springer, 2011, pp. 448-459.

[23] Zhai K, Boyd-Graber J, Asadi N, Alkhouja M. Mr. LDA: A Flexible Large Scale Topic Modeling Package using Variational Inference in MapReduce. In: 21st ACM International Conference on World Wide Web; Lyon, France; 2012. pp. 879-888.

[24] Panichella A, Dit B, Oliveto R, Di Penta M, Poshyvanyk D, De Lucia A. How to Effectively Use Topic Models for Software Engineering Tasks? An Approach Based on Genetic Algorithms. In: 2013 International Conference on Software Engineering (ICSE '13); San Francisco, CA, USA; 2013. pp. 522-531.

[25] Bagheri A, Saraee M, Jong F. ADM-LDA: An aspect detection model based on topic modelling using the structure of review sentences. Journal of Information Science 2014; 40 (5): 621-636. doi: 10.1177/0165551514538744

[26] Zheng X, Lin Z, Wang X, Lin KJ, Song M. Incorporating appraisal expression patterns into topic modeling for aspect and sentiment word identification. Knowledge-Based Systems 2014; 61: 29-47. doi: 10.1016/j.knosys.2014.02.003

[27] Wang T, Cai Y, Leung H, Lau RYK, Li Q et al. Product aspect extraction supervised with online domain knowledge. Knowledge-Based Systems 2014; 71: 86-100. doi: 10.1016/j.knosys.2014.05.018

[28] Xie W, Zhu F, Jiang J, Lim EP, Wang K. Topicsketch: Real-time bursty topic detection from twitter. IEEE Transactions on Knowledge and Data Engineering 2016; 28 (8): 2216-2229. doi: 10.1109/TKDE.2016.2556661

[29] Li C, Cheung WK, Ye Y, Zhang X, Chu D, Li X. The Author-Topic-Community model for author interest profiling and community discovery. Knowledge and Information Systems 2015; 44 (2): 359-383. doi: 10.1007/s10115-014-0764-9

[30] Liu Y, Wang J, Jiang Y. PT-LDA: a latent variable model to predict personality traits of social network users. Neurocomputing 2016; 210: 155-163. doi: 10.1016/j.neucom.2015.10.144

[31] Zoghbi Z, Vulic I, Moens MF. Latent Dirichlet allocation for linking user-generated content and e-commerce data. Information Sciences 2016; 367-368: 573-599. doi: 10.1016/j.ins.2016.05.047

[32] Yeh JF, Tan YS, Lee CH. Topic detection and tracking for conversational content by using conceptual dynamic latent Dirichlet allocation. Nerocomputing 2016; 216: 310-318. doi: 10.1016/j.neucom.2016.08.017

[33] Ekinci E, İlhan Omurca S. Concept-LDA: incorporating Babelfy into LDA for aspect extraction. Journal of Information Sciences 2019; 1: 1-20. doi: 10.1177/0165551519845854

[34] Rao Y. Contextual sSentiment topic model for adaptive social emotion classification. IEEE Intelligent Systems 2016; 31 (1): 41-47. doi: 10.1109/MIS.2015.91

[35] Xie P, Yang D, Xing EP. Incorporating word correlation knowledge into topic modeling. In: The 2015 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2015); Denver, Colorado, USA; 2015. pp. 725-734.

[36] Alam H, Ryu WJ, Lee SK. Joint multi-grain topic sentiment: modeling semantic aspects for online reviews. Information Sciences 2016; 339: 206-223. doi: 10.1016/j.ins.2016.01.013

[37] Yao L, Zhang Y, Chen Q, Qian H, Wei B et al. Mining coherent topics in documents using word embeddings and large-scale text data. Engineering Applications of Artificial Intelligence 2017; 64: 432-439. doi: 10.1016/j.engappai.2017.06.024

[38] Fu X, Sun X, Wu H, Cui L, Huang JZ. Weakly supervised topic sentiment joint model with word embeddings. Knowledge-Based Systems 2018; 147: 43-54. doi: 10.1016/j.knosys.2018.02.012

[39] Shams M, Baraani-Dastjerdi A. Enriched LDA (ELDA): combination of latent Dirichlet allocation with word co-occurrence analysis for aspect extraction. Expert Systems with Applications 2017; 80: 136-146. doi: doi.org/10.1016/j.eswa.2017.02.038

[40] Heng Y, Gao Z, Jiang Y, Chen X. Exploring hidden factors behind online food shopping from Amazon reviews: A topic mining approach. Journal of Retailing and Consumer Services 2018; 42: 161-168. doi: 10.1016/j.jretconser.2018.02.006

[41] Kandemir M, Kekeç T, Yeniterzi R. Supervising topic models with Gaussian processes. Pattern Recognition 2018; 77: 226-236. doi: 10.1016/j.patcog.2017.12.019

[42] Akın MD, Akın AA. Türk Dilleri için Açık Kaynaklı Doğal Dil İşleme Kütüphanesi: Zemberek. Elektrik Mühendisliği 2007; 431: 38-44.

[43] Manning CD, Surdeanu M, Bauer J, Finkel J, Bethard SJ et al. The stanford corenlp natural language processing toolkit. In: 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations; Baltimore, Maryland, USA; 2014. pp. 55-60.

[44] Navigli R, Ponzetto SP. BabelNet: Building a Very Large Multilingual Semantic Network. In: 48th Annual Meeting of the Association for Computational; Uppsala, Sweden; 2010. pp. 216-225.

[45] Ehrmann M, Cecconi F, Vannella D, McCrae J, Cimiano P et al. Representing Multilingual Data as Linked Data: the Case of BabelNet 2.0. In: Ninth International Conference on Language Resources and Evaluation; Reykjavik, Iceland; 2014. pp. 401-408.

[46] Wallach HM, Mimno D, McCallum A. Rethinking LDA: Why Priors Matter. In: Bengio Y, Schuurmans D, Lafferty J, Williams CKI, Culotta A (editors). Advances in Neural Information Processing Systems 22. New York, NY, USA: Curran Associates Inc., 2009, pp. 1973-1981.

[47] Chen Z, Liu B. Topic modeling using topics from many domains, lifelong learning and big data. In: 31st international conference on machine learning (ICML '14); Beijing, China; 2014. pp. 703-711.

[48] Chen Z, Liu B. Mining Topics in Documents: Standing on the Shoulders of Big Data. In: 20th ACM SIGKDD Conference on Knowledge Discovery and Data Minnig (KDD'14); New York, NY, USA; 2014. pp. 1116-1125.

[49] Mimno D, Wallach HM, Talley E, Leenders M, McCallum A. Optimizing semantic coherence in topic models. In: 2011 Conference on Empirical Methods in Natural Language Processing; Edinburgh, Scotland, UK; 2011. pp. 262-272.