Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi

Pamukkale University Journal of Engineering Sciences

An alternative word embedding approach for knowledge representation in online consumers' reviews

Çevrimiçi kullanıcı yorumlarının bilgi temsili için alternatif bir kelime gömme yaklaşımı

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

[1]Department of Computer Engineering, Faculty of Technology, Sakarya University of Applied Sciences, Sakarya, Turkey.
ekinekinci@subu.edu.tr

[2]Department of Computer Engineering, Faculty of Engineering, Kocaeli University, Kocaeli, Turkey.
silhan@kocaeli.edu.tr

Received/Geliş Tarihi: 16.02.2022 Revision/Düzeltme Tarihi: 20.06.2022 doi: 10.5505/pajes.2022.10369
Accepted/Kabul Tarihi: 19.07.2022 Research Article/Araştırma Makalesi

Abstract

Purchasing decisions in e-commerce shopping websites are highly influenced by online reviews. Although online reviews contain fine-grained consumers' opinions that reflect their preferences towards products; an important challenge, is that the number of online reviews can be very huge for fast and effective analysis. Hence, discovering the thematic structure of documents plays an important role in analyzing online reviews. The proposed system in this paper aims to discover the main consumer interests in online reviews on Turkish e-commerce websites. For this aim, a novel hybrid method combining Latent Dirichlet Allocation (LDA) and word2vec is proposed. Finally, we compare the performance of our work with those of several state-of-the-art baselines on 7 datasets collected from well-known Turkish e-commerce websites. The experimental results show how our proposed approach was able to provide significantly improved performance over baselines. Besides, our method enables us to discover very specific topics complying with consumer interests.

Keywords: Consumer reviews, Latent dirichlet allocation (LDA), Word2vec, Semantic similarity, Topic extraction.

Öz

E-ticaret alışveriş sitelerinde satın alma kararları, çevrimiçi yorumlardan oldukça etkilenir. Çevrimiçi yorumlar, ürünlere yönelik tercihleri yansıtan ayrıntılı tüketici görüşleri içerse de; önemli bir zorluk, çevrimiçi yorumların miktarının hızlı ve etkili bir analiz için çok büyük olabileceğidir. Bu nedenle, belgelerin tematik yapısını keşfetmek, çevrimiçi yorumları analiz etmede önemli bir rol oynar. Bu çalışmada önerilen sistem, Türk e-ticaret web sitelerindeki çevrimiçi yorumlardaki tüketicilerin ana ilgi alanlarını keşfetmeyi amaçlamaktadır. Bu amaçla, Gizli Dirichlet Ayırımı (GDA) ve word2vec'i birleştiren yeni bir hibrit yöntem önerilmiştir. Son olarak, çalışmamızın performansını, güncel yöntemlerin performansıyla tanınmış Türk e-ticaret sitelerinden toplanan 7 veri kümesi üzerinden karşılaştırdık. Deneysel sonuçlar, önerilen yaklaşımımızın güncel yöntemlere göre önemli ölçüde gelişmiş performans sağlayabildiğini göstermektedir. Ayrıca yöntemimiz, tüketici ilgi alanlarına uygun çok özel konuları keşfetmeyi sağlar.

Anahtar Kelimeler: Müşteri yorumları, Gizli dirichlet ayırımı (GDA), Word2vec, Anlamsal benzerlik, Konu çıkarımı.

1 Introduction

Consumer reviews are vital for e-commerce businesses. Understanding consumers' affecting factors, so that, managing consumers based on these factors is inevitable for the e-commerce world [1]. Recently, e-commerce websites which provide user textual reviews containing rich information on different products gain trust from potential consumers. This is mostly because reviews allow users to increase their awareness, evaluations, confidence, and information searches in purchasing decisions [2]. Consumer reviews contain not only the user's general opinion on products but also the consumer's fine-grained opinions towards various aspects of products. Essentially, consumer opinions are very critical as they reflect the consumer's satisfaction with aspects of a product. And also, knowing the aspects that the consumers make reviews about is very critical for the brand owners to move their brands forward. Consequently, e-commerce companies must deal with huge online review texts for success.

Understanding, organizing or modelling large collections of online reviews with unsupervised techniques are challenging problems in natural language processing (NLP) and text mining systems. When consumers evaluate online reviews about a product, and what they are interested in, it is seen that they are interested in the main themes rather than words. The thematic representations of user reviews can be efficiently used to summarize, visualize and explore the preferences of consumers. Therefore, discovering essential thematic representations of online reviews with unsupervised techniques is a key research problem for the improvement of e-commerce systems.

LDA is the most used topic model which provides the thematic representation of documents in an unsupervised manner. In LDA, the documents are modelled as a mixture of latent topics which are learnt by exploiting document-level word co-occurrences [3]. Although it is claimed that topics are semantically coherent, LDA considers only word co-occurrence distribution in the corpus, not the semantic knowledge it contains [4].

The core of our proposed work is to develop a knowledge-based system for discovering fine-grained topics that users review about a given product. Considering that in e-commerce the discovered topics from online reviews are the main influencers

[*]Corresponding author/Yazışılan Yazar

of consumer purchasing decisions and the producer sales strategies; a topic model developed for this purpose is as successful as it semantically reflects the preferences and thoughts expressed in user comments. In this study, we tried to reveal user preferences about several products which are obtained from leading product review websites in Turkey. As a consequence of these, we aimed to discover semantically coherent and meaningful aspects of online reviews with the proposed method which combines LDA with Word embeddings which is called WES-LDA.

In recent studies, word embeddings are efficiently used to represent documents as low-dimensional vectors that can capture the semantics of the documents. The embedding word vectors contain the semantic and syntactic similarity between words in the training corpus. Word2vec developed by Mikolov et al. [5] is one of the most popular methods to learn word embeddings. Therefore, in our proposed work, the word2vec method is applied to obtain a semantic representation of online reviews. Online reviews are clustered due to their word2vec representations. In every new cluster of reviews, the number of included documents tells us how much common information the cluster contains. So, the size of clusters has been included in the LDA model as a new parameter.

To summarize, the contribution of this study is as follows:

i. To represent influencers of e-commerce consumers semantically and coherently, we take advantage of the word2vec and topic model based knowledge-based system,

ii. The semantically similar documents are merged then similarity information is incorporated into the LDA as a new adaptive semantic parameter,

iii. The model is domain-independent. To prove domain independence, we introduce 7 real-world online review datasets in Turkish from different domains,

iv. We made comparisons with the state-of-the-art baselines. The experimental results indicated that our proposed method performs better in comparison with the baselines.

The rest of the paper is structured as follows. Section 2, reviews the relevant literature. Section 3 describes the proposed system in detail. The experimental study with its results is discussed in Section 4. Finally, the conclusion is presented in Section 5.

2 Background and literature review

2.1 Influence of e-commerce on consumers

E-commerce websites with reviews provided have become an important channel to make purchasing decision on a wide range of products. Therefore, determining and representing the influencers in the reviews has become very important task. More recent studies have concentrated on this point.

Yang et al. proposed to measure benefits of reviews for users by applying LDA to 35 million reviews in five categories from Amazon [6]. They evaluated results qualitatively and quantitatively and they came to this conclusion as certain aspects were not provided any extra information in terms of helpfulness. Guo et al. attempted to discover insightful topics in tourism domain which effected consumer satisfaction by means of user generated contents [7]. The experimental dataset included 266,544 user reviews for 25,670 hotels, which settled in 16 different countries. LDA extracted 19 topics to judge customer satisfaction and determine which factors were decisive in decision-making process. Heng et al. run LDA on that customer reviews on grocery products obtained from Amazon to learn factors affecting customers' online food shopping decisions [8]. The dataset contains 1,048,576 reviews from 2004 to 2014. Based on experiments, it was concluded that those four latent topics: amazon service, physical feature, subjective expression, flavor feature has impact on consumers' online shopping decision. Li et al. examined the effects of online reviews on product sales using joint sentiment-topic model which provided richer representation of user reviews [9]. The experiments were performed on 88,901 user reviews on 794 tablet computer products or SKUs. The authors extracted hedonic experience and hardware as positive topics and user interface and logistics, and customer service as negative topics from these reviews. Mou et al. performed the LDA on consumer feedbacks from cross-border ecommerce to uncover 35 topics which are play crucial role for both buyers and sellers [10]. In another study, the authors applied LDA again on consumer and firm feedbacks from cross-border e-commerce context to understand consumers' opinions to benefit both consumers and firms [11]. The dataset used in this study was multilingual such that while feedbacks of firms were in Chinese, the feedbacks of consumers were in English. They extracted thirty-five topics from each datasets and selected five most informant among them. In [12] authors proposed to discover latent dimensions in online user reviews in terms of customer satisfaction and also relationship between extracted latent factors and consumer satisfaction. In evolutionary step, 30 latent dimensions was extracted with LDA that was run on 51,110 online reviews from 1,610 different restaurants. Xu applied Latent Semantic Analysis (LSA) to reviews of the sharing economy and hotel industry to extract decision-making factors from user reviews [13]. The datasets, they used, were obtained from Airbnb for the sharing economy and Expedia for the hotel industry. For each dataset, eight influential topics and terms associated with these topics were uncovered.

With extracted topics, the effects of different topics on users were examined. Lang et al. proposed to evaluate benefits and costs of fashion renting by using real renting experiences of consumers [14]. To achieve this goal, they collected positive and negative consumer reviews from three popular fashion rental companies and applied LDA to these two datasets. While positive reviews were evaluated based on experiential value, financial value, ease of use, negative reviews were evaluated according to unsatisfied service, disappointed product performance, insufficient inventory topics.

2.2 LDA with word2vec

All existing approaches based on combination of topic model and word embeddings are quite different from ours in that topic extraction is realized by word embedding representation of documents. Further, when the literature is evaluated from another angle, it is seen that the studies differ mostly in terms of distributions. Approaches are either dirichlet based, Dirichlet Multinomial Mixture based, Gaussian based or so on with symmetric dirichlet coefficient. In WES-LDA topic extraction is realized on clusters of documents which are combined based on word embedding based similarity. On the other hand, to reflect the effect of semantic clustering, we injected the size of cluster into the model as an asymmetric dirichlet coefficient.

In [15] authors were pioneered in using LDA and word embedding together to extract semantically related topics. Since while the documents are considered as mixture of word embeddings and a topic has a multivariate Gaussian distribution over embeddings, this method has brought a different perspective to the topic models. Based on the assumption that using latent feature (LF) representation for the original features provides outperforming results in NLP studies, in [16] authors devised two new topic models for short texts. In [17]'s study three topical word embedding (TWE) models were proposed to learn different word embeddings under different topics for a word, because a word could connote different meanings under different topics. To achieve this aim LDA was implemented firstly then word embeddings were obtained based on three versions of the TWE. In TWE1, topics and words were translated to topic and word vectors separately by using skip-gram and obtained vectors were combined for topical word embeddings. TWE2 trained topical word embeddings from word-topic pairs directly with skip-gram. TWE3 used skip-gram to learn word vectors and used TWE1 to learn topic vectors. Among these three models, the most successful one was the simplest method, TWE1. Niu et al. claimed that probability was not a good way to represent features and LDA provides statistical information about co-occurrences of features in the corpus [18]. Therefore, they proposed Topic2Vec, which was the combination of neural probabilistic language model (NPLM), Word2Vec and LDA. In the model, by using LDA pre-trained topic labels were obtained as contextual information. Then these labels were included into the topics and all of them were trained with neural network language model to learn embeddings. In [19]'s paper, Gaussian Mixture Topic Model (GMTM) in which soft clustering was applied over word embedding representation of documents was proposed. GMTM assumed that Gaussian components coincide with latent topics on the embedding space and word embeddings on short text corpus were obtained with log-linear model. Wang et al. employed a hybrid method by using Word2Vec and LDA together to provide enhanced document representation. While statistical relationship between documents and topics were acquired with LDA, contextual relationship was acquired with Word2Vec [20]. Hu and Tsujii proposed a latent concept topic model (LCTM), in this model, unlike LDA, a latent topic had a probability distribution over the latent concepts rather than words and each latent concept had a Gaussian distribution over the word embeddings [21]. Li et al. carried out TopicVec by combining LDA with word embedding to improve topic coherence [22]. Moody formed lda2vec to acquire coherent topics. The proposed model learned per-document representations from embedded word, topic and document vectors by preserving semantic knowledge [23]. Word embeddings that were incorporated into the GPU-DMM was learned from external large documents. Batmanghelich et al. took advantage of von Mises-Fisher distribution (vMF) for modelling word embeddings for representation and used Hierarchical Dirichlet Process (HDP) and Stochastic Variational Inference (SVI) as a base topic model [24]. Their model was called as sHDP. With sHDP they aimed to decide topic count flexibly and explore semantic relation between embeddings. Li et al. presented Dirichlet Multinomial Mixture (DMM) and generalized P´olya urn (GPU) model based topic model (GPU-DMM) enriching with auxiliary word embeddings for short texts [25]. Indeed, besides LCTM, they offered Non-Interactive Contained LDA (nIcLDA) which was a non-interactive topic model used word embeddings as constraint to refine topics. Xun et al. formed Correlated Gaussian Topic Model (CGTM) to model topics as Gaussian distributions in the word embedding space and to learn topic correlations over continuous Gaussian topics [26]. In their models, obtained latent features were injected into LDA and one-topic-per document Dirichlet Multinomial Mixture (DMM) to improve the topic-to-word mapping. In [27] the authors developed two document-enriching methods for extracting topics from short texts. The first method, which was called as Co-Frequency Expansion (CoFE), assumed high co-frequent words had high probability under the same topic. The second method Distributed Representation-based Expansion (DREx) used word embeddings to expand documents. Law et al. suggested Latent Topical Skip-Gram (LTSG) which was based on polysemous word embeddings [28]. Polysemous-word models and topic models were learnt mutually in this model. While polysemous-word embeddings were obtained from word embeddings and global topic embeddings, topic word distribution was learnt by using polysemous words. Experimental results showed that LTSG outperformed state-of-art methods on text classification and topic mining. Garcia-Pablos et al. proposed an almost unsupervised model, W2VLDA, in which topic modelling was combined with continuous word embeddings and a Maximum Entropy classifier [29]. Zhao et al. enacted a novel topic model that customized a multi-layer Gamma Belief Network for deeply representation of topics to reveal inter-topic and intra-topic structures [30]. Shi et al. proposed topic models, semantics-assisted non-negative matrix factorization (SeaNMF) and sparse SeaNMF (SSeaNMF) were solved with block coordinate descent algorithm [31]. For this, the semantic knowledge, which included relationship between the words and their contexts, was extracted by using a skip-gram. Viegas et al. presented a new document representation with pre-trained word embedding for non-probabilistic matrix factorization (NMF) for topic modelling [32]. The proposed model called as Cluster of Words (cluWords) used word embeddings to enhance document representation with semantic and syntactic relationships according to neighbor relationships among words. Besides, importance of obtained cluWords was measured with TF-IDF scheme to represent topics. Li et al. devised Topic Modeling and Sparse Autoencoder (TMSA) a mutual learning mechanism by unifying topic models and word embeddings to improve interpretability of extracted topics [33]. In TMSA, autoencoder with back propagation mechanism was used to feedback word embeddings from global topics and local context. Al-Janabi et al. proposed to improve approximation of conditional distribution using word embedding based semantic similarity for semantically related topics [34].

3 Proposed method

3.1 Pre-processing

The main goal of the pre-processing is to improve the quality of data by transforming text from human language to machine-readable language. The main characteristic of user reviews is electronic word-of-mouth behavior. Therefore, white spaces, numbers, symbols, emoticons, capital letters, character repetitions, typo errors, abbreviations commonly occurred in these type of documents. These type of usages should be handled so seriously for accurate analysis of documents [35]. For this reason, noisy labels such as urls, white spaces, numbers, symbols, emoticons are removed from documents; all letters are converted to lower case because of case sensitivity; abbreviations are expanded. Character repetitions and typo errors are corrected with Zemberek-Turkish Natural Language Processing Library [36]. Punctuations and diacritics, that are considered useless, are also removed from documents.

Stopwords are quite often occurring, informative words that appear in documents. The only task is to assist in the creation of the documents; stopwords are also removed from them. Out of these pre-processing steps, stemming is applied to the documents for reducing words to their base or root form.

3.2 Extraction of word embeddings

Learning what a word means is very crucial for natural language processing studies and provides the addressing of the word-sense ambiguity problem in a broader and deeper way. In recent years, word embedding has emerged as an important method used in document representation. Since, for a word, Word embedding can capture the context of this word in the document that the word refers to and can explore the semantic and syntactic relation associated with the other words in the document. It is well-documented that the similarity of word embeddings reflects many different relations between the given words, e.g. synonymy, antonym, hypernym or hyponymy [37].

The main purpose of WES-LDA is to strengthen the probability distribution from LDA by incorporating semantically similar documents into the model. In order to obtain semantic similarity, we model documents semantically by representing each word with its context vector obtained with skip-gram.

Skip-gram which is an n-gram based model is a learning model in Word2vec [38]. Skip-gram model realized a neural network (NN) architecture and has three layers namely input, projection, and output. By using this architecture, the Skip-gram model takes a particular word as input and predicts the context words as output.

The reason why we use skip-gram is that the context of the document can be captured independently of n-grams for similar documents [39]. In this study, we utilized pre-trained word embedding which was trained on Wikipedia pages for Turkish.

3.3 TF-IDF matrix calculation

In the remaining steps, the documents are analyzed according to a term frequency inverse document frequency (TF-IDF) matrix which is generated based on the word embedding vectors. The similar documents are clustered by using the TF-IDF matrix.

Although TF-IDF is an old method, it is still one of the most commonly used vectorization algorithm for text documents. TF-IDF regards two documents as similar if these two documents share not only uncommon but also distinctive words. While the distinctiveness of a word is directly proportional with number of passes over the relevant documents, is inversely proportional with number of passes over the whole collection.

3.4 Similarity calculation and knowledge extraction

In order to improve LDA, it is intended to be: documents having semantic relationship with each other should be in the same cluster together and the knowledge obtained from these Semantic Co-occurence Clusters is used for improvement of the LDA. In the proposed model, semantic similarity is calculated over word embeddings however combining is applied over original documents. By combining similar documents together in the same cluster, with the idea that semantically similar documents are mentioning about the same topics, so semantically related words together and co-occurrence relation of LDA strengthen with the semantic similarity. If we explain this through user reviews in the corpora, we can take these two words “tat” and “lezzet”. These two words are not in the same user reviews in the corpora. However, for these two words word embedding vectors include both “lezzet” and “tat” together, so the sentences include these two are combined, and the likelihood of these two words appearing under the same topic is increasing. Of course, an optimization criterion for putting documents in the same cluster should be applied. Similarity greater than 0.5 is accepted as optimization criterion for merging two similar documents.

3.5 Merging documents

For each Semantic Co-occurrence cluster, the documents that make up the cluster are merged and one document from each cluster is obtained. Based on the clusters, number of documents in the corpora is decreased.

3.5.1 Topic extraction

LDA is one of the most prevalent generative probabilistic topic models, discovers the latent topics with relative words for each topic from consumer reviews [40]. Because of the idea that priors have no significant effect on the LDA, in almost all studies in the literature symmetric Dirichlet priors are used to extract topics. However, in this study, asymmetric Dirichlet priors are used to learn document topic distribution from user reviews. There are two reasons for this. One is that it is demonstrated that considering asymmetric Dirichlet priors to learn document topic distribution gives the best results in terms of model performance [41]. The other is that two documents are semantically similar, and the topic distributions of documents cannot be assumed to be independently sampled. Nevertheless, in this study, semantically similar documents are combined into a single document so; there is no significant similarity between the new documents composed. Accordingly, the topic distributions of documents can be sampled independently and asymmetric Dirichlet priors are used to learn document topic distribution in WES-LDA. The generative process of WES-LDA for each merged consumer review is described as follows:

Algorithm 1: The generative process of WES-LDA

```
1: for each topic $k$, $k \in [1,K]$ do
2:     Determine the distribution of words within topics: $\phi_k \sim$ Dirichlet($\beta$)
3: end for
4: for each merged consumer reviews $g$, $g \in [1,G]$ do
5:     Determine the distribution of topics within consumer reviews:
6:     $\theta_g \sim$ Dirichlet($\alpha u$)
7:     for each word $w_c$ in consumer review $g$ do
8:         Randomly choose a topic from the distribution over topics in step 5:
9:         $z_{g,c} \sim$ Mult($\theta_g$)
10:        Randomly choose a word from the step 2: $w_{g,c} \sim$ Mult($\phi_k$, $z_{g,c}$)
11:    end for
12: end for
```

For the model introduced the conditional posterior probability of topic k in document g given z as:

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

In the Equation (1), p($z_{g,c}$ = k | $z_{\neg g,c}$, αu) is the probability of the assignment of current word c_g in consumer review g to each topic, conditioned other than all other words $c_{\neg g}$. In the denominator 1 is subtracted from Cg number of words in document g due to the current word is not included. From Equation (1), full conditional distribution p($z_{g,c}$ = k | $z_{\neg g,c}$,w) is calculated as follows:

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

In Equation (2), second ratio indicates the probability of topic k in document g. $c_{w,k}$ is the number of times word w is assigned to topic k in the whole consumer review collection. The number of times topic k is used in the whole consumer review collection but not including the current word is $c_{w',k}$.

4 Experimental study

This section describes the experiments conducted with the purpose of extracting semantically related topics from user reviews. Each performance analysis is carried out with seven different Turkish datasets, which are first used in this study. The evaluation metrics computed for all the experiments in this work are topic coherence and TF-IDF coherence, for interpretability of topics' semantics. In addition, we put forth extracted topics to show semantic relation and the ability to capture details. The experiments are conducted in Netbeans IDE 8.1 for Windows using the Java language.

4.1 Datasets

To evaluate the ability of topic extraction which is achieved by WES-LDA, we use 7 different datasets in Turkish which have not been previously used in any study. The datasets are in different domains, and they represent different aspects of real-world user reviews. The hotel dataset is received from a known website otelpuan.com. The remaining six datasets, which are tea machine, headphones, modem, mobile phone, television and USB in electronic domains, are from the hepsiburada.com, a popular e-shopping website.

At the stage of collecting user reviews from websites, Konstanz Information Miner (KNIME) is used. KNIME is an open-source data analysis, reporting and aggregation tool, which allows data to be understood and workflows for data processing [42]. KNIME provides its user components for financial data analysis, text processing, time series, visualization, and so on. Besides all these features, it offers great convenience and simplicity for applications; therefore, KNIME is preferred in this study.

Based on the [43]'s study, in WES-LDA, merged documents based on the similarity of word embeddings are the input of the LDA. It is important to select an appropriate threshold value for an accurate clustering to model semantic similarity accurately. That's why, all possible threshold values are examined and 0.5 is selected as the threshold value. If the similarity between the most similar two documents is greater than 0.5, these two documents are placed in the same cluster. In the document collections, while the maximum number of documents in the clusters varies, the minimum number of documents in sets is equal to one for each collection. After all clusters are formed, the documents in every cluster are merged as a new document.

Number of reviews for original and similarity-based datasets are introduced, and some statistics are reported in Table 1.

4.2 Parameter settings

All experiments are run with 50, 100, 200, 500 and 1000 iterations of Gibbs sampling and Dirichlet hyperparameters α and β are set to 50/K and 0.001 respectively. These are the common values in topic modelling studies. While the β value is symmetric, the α value becomes asymmetric by injecting u into the LDA module. For each input document, u is the number of the merged document to comprise this document. K, the number of extracted topics, is determined as 100 and every topic is published with the first ten words. The same parameter values are utilized for baselines to compare with WES-LDA quantitatively and qualitatively.

4.3 Compared models

We compare our model with six LDA models:

- ✓ LDA: LDA is the most widely used simplest generative topic model and no external knowledge regarding the relationship between words is used to extract topics [44, 45],
- ✓ TF-IDF LDA: We devised TF-IDF to compare the semantic representation of the documents and the original representation of the documents in terms of semantic coherence and performance. The only difference between the models is the document representation used in calculating semantic similarity. As in WES-LDA, similar documents are merged to compose a new document and element count u in a cluster is added to LDA. While merging document for TF-IDF LDA 0.5 is selected as threshold like in WES-LDA. After merging documents, the document counts are decreases to 1540 for hotel, 367 for tea machine, 343 for headphones, 362 for modem, 1568 for mobile phone, 438 for television and 801 for USB,

Table 1. Statistics of the datasets.

Dataset	#Reviews (Original)	Average #Word (Original)	#Reviews (Similarity)	Average #Word (Similarity)	#Max. Reviews in the Clusters
Hotel	10540	7	1588	46.45	162
Tea machine	2485	5.93	386	38.17	149
Headphones	2169	5.14	376	33.94	47
Modem	2460	6.89	404	41.97	100
Mobile phone	9409	6.1	1597	35.94	467
Television	2931	8.78	476	54.05	76
Usb	5455	4.15	832	27.24	636

- ✓ Dirichlet Multinomial Mixture (DMM): DMM models each document as mixture of unigrams [46]. DMM is EM based and takes into account the assumption that each document contains only one subject,
- ✓ Biterm Topic Model (BTM): BTM extracts topics from documents by learning the generation of word co-occurrence patterns (biterms) directly to improve topic coherence [47],
- ✓ Lifelong Topic Model (LTM): LTM is a lifelong topic model which injects must-links learned automatically from documents as prior knowledge to obtain coherent topics [48],
- ✓ topic modeling with Automatically generated Must-links and Cannot-links (AMC): AMC is the improved version of LTM with automatically extracted cannot-links to improve quality of topics in terms of topic coherence [49].

4.4 Evaluation metrics

Our aim with this study is to extract semantically related topics from user reviews. Therefore, we compare WES-LDA semantically with LDA, TF-IDF LDA, DMM, BTM, LTM and AMC. Although perplexity is a conservative model to evaluate topic models by using held-out test data, it is not able to measure interpretability of topics based on human judgement [4]. For this reason, topic coherence and TF-IDF coherence, which are compatible with human judgements about semantic relation, coherency and meaningfulness of topics, are preferred in this study [40].

Topic coherence measures the relevance of the extracted topics to the decisions of human. Topic coherence is computed as in Equation (3).

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

$V(k) = (v^{1}{}_{k}, v^{2}{}_{k}, \ldots, v^{1}{}_{S})$ represents the list of top S words in the topic k. $D(v_n{}^{(k)}; v_l{}^{(k)})$ is co-document frequency of word types v_n and v_l. 1 in the nominator is used as smoothing. $D(v_l{}^{(k)})$ in the denominator is the document frequency of word type v_l. The higher value of the topic coherence, the greater the coherence of the extracted topics in themselves.

Although topic coherence is a successful measure, it also includes some disadvantages. Topic coherence cannot capable of distinguishing between high frequency and informative words which compose the topic. To overcome this disadvantage, TF-IDF coherence, modified version of topic coherence, is proposed [50]. TF-IDF coherence is computed as shown in Equation (4).

$$c_{tf-idf}(t,W_t) = \sum_{w_1\neq w_2,w_2\in W_t} log\frac{\sum_{d:w_1,w_2\in d} tf-idf(w_1,d)tf-idf(w_2,d)+\epsilon}{\sum_{d:w_1\in d} tf-idf(w_1,d)} \quad (4)$$

In the Equation (4) W_t is the list of top words in topic t and tf–idf(w_1, d) is the TF-IDF score of the word w_1 in W_t. For a word w in document d tf–idf(w, d) score is calculated by the Equation (5).

$$tf-idf(w,d)=tf(w,d)idf(w) = \left(\frac{1}{2}+\frac{f(w,d)}{max_{w'\in d}f(w',d)}\right)log\left(\frac{|D|}{|d'\in D: w\in D'|}\right) \quad (5)$$

In the equation above, f(w, d) is the number of occurrence of word w in document d, maxf(w′, d) is the number of maximum occurrence in document d. And the right side of the equation is the logarithm of average number of words per document.

4.5 Evaluation results

In WES-LDA, the input documents of the model are combined by including word embedding based similarity knowledge. This knowledge is used to extract more coherent topics and meaningful topics from user reviews. To assess the qualitative and quantitative performance of the proposed model, datasets with seven different domains in Turkish are selected. Average topic and TF-IDF coherence of top 10 words with 100 topics is calculated over original datasets for LDA, DMM, BTM, LTM and AMC and is calculated over merged datasets for WES-LDA and TF-IDF LDA.

Figure 1 shows the average topic coherence on each domain for 50, 100, 200, 500 and 1000 iteration counts respectively. Based on the results, we can say that merging documents based on similarity achieves high performance in terms of topic coherence. Even runs for 50 iterations, WES-LDA and TF-IDF LDA are more successful than 1000 iterations of LDA, DMM, BTM, LTM and AMC. Between WES-LDA and TF-IDF LDA, WES-LDA is more successful. In addition, when the methods are examined according to increasing iteration counts, WES-LDA becomes more successful as the number of iterations increased.

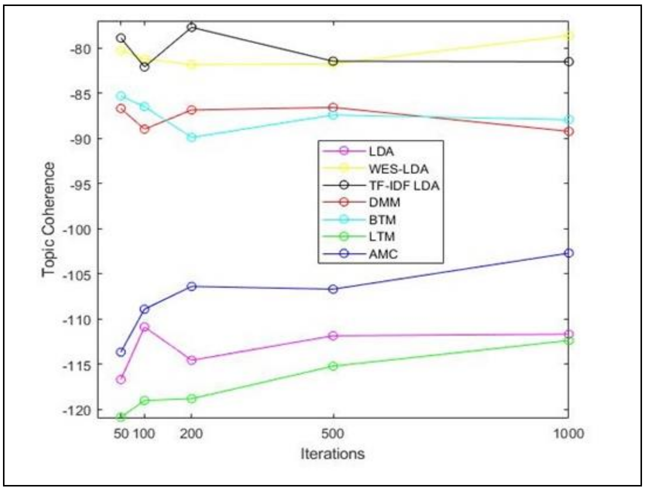

Figure 1. Topic coherence.

Figure 2 shows the average TF-IDF coherence on each domain for 50, 100, 200, 500 and 1000 iteration counts respectively. This evaluation metric confirms that the similarity based merging achieves high performance and also confirms the results obtained with the topic coherence. Comparing tf-id coherence with topic coherence, the only difference is that TF-IDF LDA is more successful than WES-LDA. However, it is clear that both methods have similar successes in terms of coherence and are quite successful compared to the other five methods.

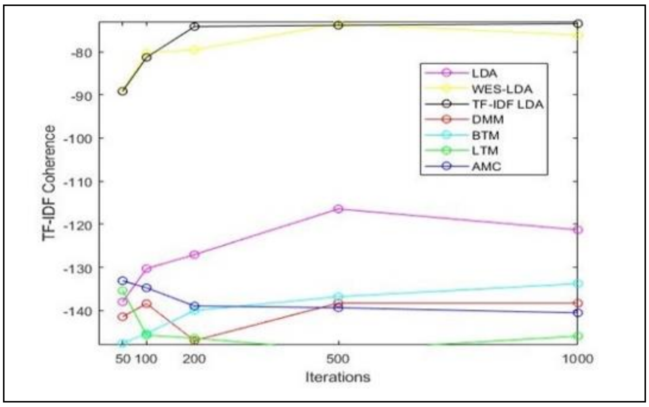

Figure 2. TF-IDF coherence.

In order to better understanding of consumer reviews in terms of factors affecting the consumers, we select top topic words over the topics obtained with WES-LDA. For randomly selected four consumer dataset top 5 topics with relative top 5 words are represented in Figure 3. The reason why we choose 5 topics among all is these topics fully reflecting the consumers' satisfaction. Given that consumer reviews are in Turkish, to better help the understanding the topics, we translated them into English. By using most likely words in the topics the raised topics are labeled manually

In order to make a qualitative analysis, the seven tables are examined separately. For instance, the top 5 topics for headphones dataset are acoustic, details, noise, performance, design. The topics and related words raised are very important for both potential consumers and companies in terms of expressing the consumers' satisfaction or dissatisfaction with these topics. While the consumers can make decision based on these issues, companies can make improvements regarding the specified topics, if necessary. If performance is important for the consumer who wants to buy headphones, he will decide accordingly. If the reviews about the performance are negative, he will not buy the product and companies can improve the performance, or vice versa. Consequently, we make a generalization on the topics obtained:

- ✓ Good performance affects awareness of consumers on products,
- ✓ Good performance affects information searches of consumers on products,
- ✓ Good performance affects evaluation of consumers on products,
- ✓ Good performance affects confidence of consumers on products.

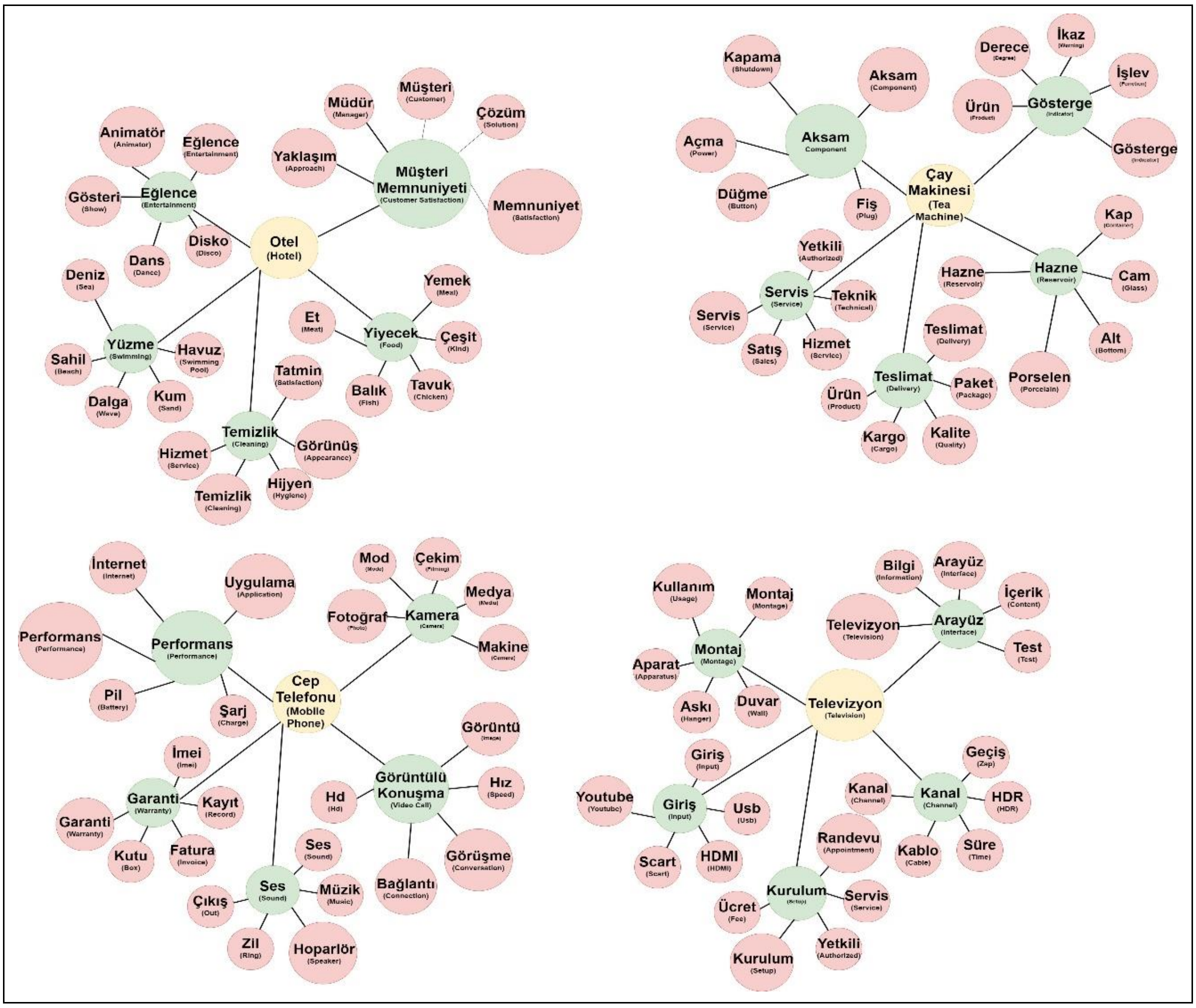

Figure 3. Extracted topics.

In addition to all these evaluations, topics extracted with WES-LDA are compared with subjects from other 6 baselines. Comparative results are reported only for headphones in Table 2 due to space shortage. Errors are italicized in red.

The comparative results show that topics obtained with WES-LDA are semantically related, coherent, meaningful, and easily labeled; however, the baseline results are of quite low quality. Consequently, when we compare the WES-LDA with baselines for the headphones domain; it is obvious that the topic words of WES-LDA provide a predominant representation of the headphones domain.

5 Conclusions

Managing consumer reviews is very important for both consumers and companies and is a very important subject of study. To accomplish this goal, this study intends to investigate the factors affecting consumers' decisions in Turkish e-commerce websites for different domains by using WES-LDA. Unlike other studies that only analyzed one domain, we used consumer reviews from seven different domains not only to extract the affecting factors but also to demonstrate how successful the WES-LDA is for each domain. In addition, this paper is the first to determine factors affecting consumers with LDA based model from Turkish e-commerce websites.

This empirical study shows that companies can determine which factors affect consumers' awareness, information searches, evaluations, and confidence on products easily. Using knowledge obtained from WES-LDA to better understand the consumer can help the development of the product according to the wishes of the consumers. Further, this paper shows that extracted topics can help potential consumers make an informed decision during the purchasing process.

Since the LDA model only analyzes the co-occurrence distribution of words in the sentence, not the semantic information it contains, semantically relevant, coherent and meaningful topics cannot be generated using the LDA model alone. This is the main disadvantage of LDA. To address this defect, a new semantic LDA model is proposed in this research. The proposed approach to do this involves a well-known semantic measure derived from a set of semantically comparable documents to influence the document-topic distribution. Instead of words, the feature space of documents is created using word embeddings to obtain semantically similar documents. Using these units to define the feature space ensures that semantic information is appropriately included in the model. Experimental results of the WES-LDA model show that injecting semantic similarity into the model results in topics that are semantically related, consistent, and relevant. In the future, we will evaluate our method in more languages with different word-placement algorithms to explore differences. We also plan to apply WES-LDA to multilingual document collections.

6 Author contribution statements

In the work carried out, the Ekin EKİNCİ contributed to forming the idea, making the design and literature review, collecting data, performing the analysis, and examining the results; The Sevinç İLHAN OMURCA contributed to the creation of ideas, making the design, checking the spelling and checking the article in terms of content.

7 Ethics committee approval and conflict of interest statement

"There is no need to obtain permission from the ethics committee for the article prepared."

"There is no conflict of interest with any person/institution in the article prepared."

8 References

[1] Zhu G, Wu Z, Wang Y, Cao S, Cao J. "Online purchase decisions for tourism e-commerce". *Electronic Commerce Research and Applications,* 38, 1-13, 2019.

[2] Huang Z, Benyoucef M. "The effects of social commerce esign on consumer purchase decision-making: An empirical study". *Electronic Commerce Research and Applications*, 25, 40-58, 2017.

[3] Griffiths TL, Steyvers M. *Prediction and semantic association*. Editors: Thrun S, Saul LK, Schölkopf B. Advances in neural information processing systems 16, Proceedings of the Seventeenth Annual Conference on Neural Information Processing Systems, 11-18, Vancouver, Canada, 2003.

[4] Chang J, Gerrish S, Wang C, Boyd-Graber JL, Blei DM. *Reading Tea Leaves: How Humans İnterpret Topic Models*. Editors: Bengio Y, Schuurmans D, Lafferty J, Williams C, Culotta A. Advances in Neural İnformation Processing Systems 22, Proceedings of the Twenty-Third Annual Conference on Neural Information Processing Systems, 288-296, Vancouver, Canada, 2009.

[5] Mikolov T, Chen K, Corrado G, Dean J. "Efficient estimation of word representations in vector space". *arXiv*, 2022. https://arxiv.org/abs/1301.3781.

[6] Yang Y, Chen C, Bao FS. "Aspect-based helpfulness prediction for online product reviews". *2016 IEEE 28th International Conference on Tools with Artificial Intelligence*, San Jose, CA, USA, 6-8 November 2016.

Table 2. Topic words for headphone domain about acoustic.

WES-LDA	TF-IDF LDA	LDA	AMC	BTM	DMM	LTM
kulaklık	ses	seviye	ses	akustik	*teslim*	ses
akustik	*yıldız*	kaynak	müzik	kulak	akustik	ayar
stereo	ürün	*seçenek*	kulaklık	ürün	ürün	cızırtı
ses	enstrüman	*telefon*	ürün	*fiyat*	*fiyat*	performans
yalıtım	elektronik	akustik	net	ses	ses	*boyut*
ürün	akustik	*üretim*	kulak	jbl	jbl	kulaklık
radyo	şarkı	*kusur*	*fiyat*	kulaklık	kulaklık	*garanti*
aksesuar	*zevk*	efekt	*silikon*	mikrofon	mikrofon	kulak
müzik	pürüz	*kategori*	*metal*	*plastik*	*plastik*	kalite
mikrofon	kulak	ürün	performans	parça	parça	*boy*

[7] Guo Y, Barnes SJ, Jia Q. "Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation An empirical study". *Tourism Management*, 59, 467-483, 2017.

[8] 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*, 42, 161-168, 2018.

[9] Li X, Wu C, Mai F. "The effect of online reviews on product sales: A joint sentiment-topic analysis.". *Information & Management*, 56, 172-184, 2019.

[10] Mou J, Ren G, Qin C, Kurcz K. "An exploration of cross-border e-commerce consumer feedbacks: An LDA approach". *The Seventeenth Wuhan International Conference On E-Business*, Wuhan, P.R. China, 25-27 May 2018.

[11] Mou J, Ren G, Qin C, Kurcz K. "Understanding the topics of export cross-border e-commerce consumers feedback: an LDA approach". *Electronic Commerce Research*, 19, 749-777, 2019.

[12] Situmeang F, de Boer N, Zhang A. "Looking beyond the stars: A description of text mining technique to extract latent dimensions from online product reviews.". *International Journal of Market Research*, 62, 195-215, 2020.

[13] Xu X. "How do consumers in the sharing economy value sharing? Evidence from online reviews". *Decision Support Systems*, 128, 1-13, 2020.

[14] Lang C, Li M, Zhao L. "Understanding consumers' online fashion renting experiences: A text-mining approach". *Sustainable Production and Consumption*, 21, 132-144, 2020.

[15] Das R, Zaheer M, Dyer C. *Gaussian lda for topic models with Word embedding*. Editors: Zong C, Strube M. the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, 795-804, Beijing, China, 2015.

[16] Nguyen DQ, Billingsley R, Du L, Johnson M. "Improving topic models with latent feature word representations". *Transactions of the Association for Computational Linguistics*, 3, 299-313, 2015.

[17] Liu Y, Liu Z, Chua TS, Sun M. "Topical word embeddings". *Twenty-Ninth AAAI Conference on Artificial Intelligence*, Austin, Texas, USA, 25-30 January 2015.

[18] Niu L, Dai X, Zhang J, Chen J. "Topic2Vec: learning distributed representations of topics". *2015 International Conference on Asian Language Processing*, Suzhou, China, 24-25 October 2015.

[19] Sridhar VKR. "Unsupervised topic modeling for short texts using distributed representations of words". *1st Workshop on Vector Space Modeling for Natural Language Processing*, Denver, Colorado, USA, 5 June 2015.

[20] Wang Z, Ma L, Zhang Y. "A hybrid document feature extraction method using latent Dirichlet allocation and word2vec". *2016 IEEE First International Conference on Data Science in Cyberspace*, Changsha, China, June 2016.

[21] Hu W, Tsujii J. *A Latent Concept Topic Model for Robust Topic Inference Using Word Embeddings*. Editors: Erk K, Smith NA. 54th Annual Meeting of the Association for Computational Linguistics, 380-386 Berlin, Germany, 2016.

[22] Li S, Chua TS, Zhu J, Miao C. *Generative Topic Embedding: A Continuous Representation of Documents*. Editors: Erk K, Smith NA. 54th Annual Meeting of the Association for Computational Linguistics, 666-675, Berlin, Germany, 2016.

[23] Moody CE. "Mixing dirichlet topic models and word embeddings to make lda2vec". *arXiv*, 2022. https://arxiv.org/abs/1605.02019

[24] Batmanghelich K, Saeedi A, Narasimhan K, Gershman S. *Nonparametric Spherical Topic Modeling With Word Embeddings*. Editors: Erk K, Smith NA. 54th Annual Meeting of the Association for Computational Linguistics, 537-542, Berlin, Germany, 2016.

[25] Li C, Wang H, Zhang Z, Sun A, Ma Z. "Topic modeling for short texts with auxiliary word embeddings". *39th International ACM SIGIR conference on Research and Development in Information Retrieval*, Pisa, Italy, 17-21 July 2016.

[26] Xun G, Li Y, Zhao WX, Gao J, Zhang A. "A correlated topic model using word embeddings". *Twenty-Sixth International Joint Conference on Artificial Intelligence*, Melbourne, Canada, 19-25 August 2017.

[27] Bicalho P, Pita M, Pedrosa G, Lacerda A, Pappa GL. "A general framework to expand short text for topic modeling". *Information Sciences*, 393, 66-81, 2017.

[28] Law J, Zhuo HH, He J, Rong E. "Ltsg: Latent topical skipgram for mutually learning topic model and vector representations". *arXiv,* 2022. https://arxiv.org/abs/1702.07117

[29] Garc´ıa-Pablos A, Cuadros M, Rigau G. "W2VLDA: almost unsupervised system for aspect based sentiment analysis". *Expert Systems with Applications*, 91, 127-137, 2018.

[30] Zhao H, Du L, Buntine W, Zhou M. *Inter and Intra Topic Structure Learning with Word Embeddings*. Editors: Dy j, Krause A. 35th International Conference on Machine Learning, 5892-5901, Stockholm, Sweden, July 2018.

[31] Shi T, Kang K, Choo J, Reddy CK. "Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations". *2018 World Wide Web Conference*, Lyon, France, 23-27 April 2018.

[32] Viegas F, Canuto S, Gomes C, Luiz W, Rosa T, Ribas S, Rocha L, Gonçalves MA. "CluWords: Exploiting Semantic Word Clustering Representation for Enhanced Topic Modeling". *Twelfth ACM International Conference on Web Search and Data Mining,* Melbourne, Canada, 11-15 February 2019.

[33] Li D, Zhang J, Li P. *TMSA: A Mutual Learning Model for Topic Discovery and Word Embedding*. Editors: Berger -Wolf T, Chawla N. 2019 SIAM International Conference on Data Mining, 684-692, Calgary, Alberta, Canada, 2019.

[34] Al-Janabi OM, Malim NHAH, Cheah YN. *Aspect Categorization Using Domain-Trained Word Embedding and Topic Modelling*. Editors: Zakaria Z, Ahmad R. Advances in Electronics Engineering, 191-198, Kuala Lumpur, Malaysia, December 2019.

[35] Çoban Ö, Tümüklü Özyer G. "The impact of term weighting method on Twitter sentiment analysis". *Pamukkale University Journal of Engineering Sciences*, 24(2), 283-291, 2018.

[36] Akın MD, Akın AA. "Türk dilleri için açık kaynaklı doğal dil işleme kütüphanesi: ZEMBEREK". *Elektrik Mühendisliği*, 431, 38-44, 2007.

[37] Mykowiecka A, Marciniak M, Rychlik P. “Testing word embeddings for Polish”. *Cognitive Studies*, 17, 1-19, 2017.

[38] Ekinci E. “Classification of Imbalanced Offensive Dataset - Sentence Generation for Minority Class with LSTM”. *Sakarya University Journal of Computer and Information Sciences*, 5(1), 121-133, 2022.

[39] Guthrie D, Allison B, Liu W, Guthrie L, Wilks Y. “A closer look at skip-gram modelling”. *Fifth International Conference on Language Resources and Evaluation*, Genoa, Italy, 24-26 May 2006.

[40] Ekinci E, Ilhan Omurca S. “Concept-LDA: Incorporating Babelfy into LDA for aspect extraction”. *Journal of Information Science*, 46, 406-418, 2020.

[41] Wallach HM, Mimno DM, McCallum A. *Rethinking LDA: Why Priors Matter.* Editors: Bengio Y, Schuurmans D, Lafferty J, Williams C, Culotta A. Twenty-Third Annual Conference on Neural Information Processing Systems, 1973-1981, Vancouver, Canada, 2009.

[42] Atıcı B, Ilhan Omurca S, Ekinci E. “Product aspect detection in customer complaints by using latent dirichlet allocation”. *2017 International Conference on Computer Science and Engineering,* Antalya, Türkiye, 5-8 October 2017.

[43] Ekinci E, Ilhan Omurca S. “NET-LDA: a novel topic modeling method based on semantic document similarity”. *Turkish Journal of Electrical Engineering & Computer Sciences*, 28, 2244-2260, 2020.

[44] Blei DM, Ng AY, Jordan MI. “Latent dirichlet allocation”. *Journal of Machine Learning Research*, 3, 993-1022, 2003.

[45] Salur MU, Aydın İ, Jamous M. “An ensemble approach for aspect term extraction in Turkish texts”. *Pamukkale University Journal of Engineering Sciences*, 28(5), 769-776 2022.

[46] Nguyen DQ. “jLDADMM: A Java package for the LDA and DMM topic models”. *arXiv*, 2022. https://arxiv.org/abs/1808.03835

[47] Yan X, Guo J, Lan Y, Cheng X. “A biterm topic model for short texts”. *22nd International Conference on World Wide Web*, Rio de Janeiro, Brasil, 13-17May 2013.

[48] Chen Z, Liu B. “Topic modeling using topics from many domains, lifelong learning and big data”. *31st International Conference on Machine Learning*, Beijing, China, 21-26 June 2014.

[49] Chen Z, Liu B. “Mining topics in documents: standing on the shoulders of big data”. 2*0th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, USA, 24-27 August 2014.

[50] Nikolenko SI, Koltcov S, Koltsova O. “Topic modelling for qualitative studies”. *Journal of Information Science*, 43, 88-102, 2017.